SQuARM-SGD: Communication-Efficient Momentum SGD for Decentralized Optimization

Navjot Singh, *Student Member, IEEE*, Deepesh Data, Jemin George, and Suhas Diggavi, *Fellow, IEEE*

***Abstract*—In this paper, we propose and analyze SQuARM-SGD, a communication-efficient algorithm for decentralized training of large-scale machine learning models over a network. In SQuARM-SGD, each node performs a fixed number of local SGD steps using Nesterov's momentum and then sends sparsified and quantized updates to its neighbors regulated by a locally computable triggering criterion. We provide convergence guarantees of our algorithm for general (non-convex) and convex smooth objectives, which, to the best of our knowledge, is the first theoretical analysis for compressed decentralized SGD with momentum updates. We show that the convergence rate of SQuARM-SGD matches that of vanilla SGD. We empirically show that including momentum updates in SQuARM-SGD can lead to better test performance than the current state-of-the-art which does not consider momentum updates.**

***Index Terms*—Decentralized optimization; communication efficiency; Nesterov momentum.**

I. Introduction

As machine learning gets deployed over edge (wireless) devices (in contrast to datacenter applications), the problem of building learning models on local (heterogeneous) data with communication-efficient training becomes important. These applications motivate learning when data is collected/available locally, but devices collectively help build a model through wireless links with significant communication rate (bandwidth) constraints.[1] Several methods have been developed recently to obtain communication-efficiency in *distributed* stochastic gradient descent (SGD). These methods can be broadly divided into two categories. In the first one, workers *compress* information/gradients before communicating - either with *sparsification* [2]–[6], *quantization* [7]–[11], or both [12]. Another way to reduce communication is to skip communication rounds while performing a certain number of *local SGD* steps, thus trading-off computation and communication time [13]–[15]. Since momentum-based methods generally converge faster and generalize well, they have been adopted ubiquitously for training large-scale machine learning models [16].

To reduce communication load on the central-coordinator in the distributed framework, a *decentralized* setting has been considered in literature [17], where the central coordinator is absent, and training is performed collaboratively among workers, which are connected by a (sparse) graph.[2] Compressed communication has been studied recently for decentralized training as well [18]–[22]. Out of these [18], [20]–[22] only employ either quantization or sparsification (without local iterations or event-triggered communication), whereas, [19] also incorporates event-triggering to achieve communication efficiency; see related work for a detailed comparison. We would like to remark two important aspects of these works: (i) They rely on strong set of assumptions for their theoretical analyses: all of them assume a uniform bound on variance of stochastic gradients and also on the gradient dissimilarity across the clients, while [19]–[22] assume a bound on the second moment of stochastic gradients. (ii) None of these works incorporates momentum in their theoretical analyses, which has been very successful in achieving good generalization error in training large-scale machine learning models.

In this paper, we propose and analyze SQuARM-SGD,[3] a communication efficient SGD algorithm for decentralized optimization that incorporates Nesterov's momentum, compression and local iterations while considering a much weaker set of assumptions than existing literature.

For compression, SQuARM-SGD uses both sparsification and quantization. For event-triggered communication, each worker first performs a certain number of *local SGD* iterations with momentum updates; then in order to further reduce communication, it only does so if there is a significant change in the local model parameters (greater than a prescribed threshold) since its last communication. If there is a significant model change, the worker communicates a sparsified and quantized version of (the difference of) its local parameters (model) to its neighbors. Therefore, this combines lazy updates along with quantization and sparsification to enable communication-efficient decentralized training.

Our contributions. In this paper, we propose and analyze SQuARM-SGD, a communication efficient decentralized training algorithm incorporating compression and local iterations. Our analysis is the first to establish convergence rates of compressed decentralized training algorithms with momentum. We provide separate convergence results for SQuARM-SGD with two sets of assumptions: (i) Commonly used assumptions in decentralized optimization, including bounded second mo-

This work was partially supported by NSF grants #2007714, #1955632, by UC-NL grant LFR-18-548554 and by Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196.

A preliminary version of this work has appeared in the *IEEE International Symposium on Information Theory* (ISIT) 2021.

Navjot Singh, Deepesh Data, and Suhas Diggavi are with the University of California, Los Angeles, CA 90095 USA (e-mail: navjotsingh@ucla.edu; deepesh.data@gmail.com; suhas@ee.ucla.edu). Jemin George is with the US Army Research Lab, Adelphi, MD 20783 USA (e-mail: jemin.george.civ@mail.mil).

[1]This is also motivated by federated learning [1], which is studied mostly for the client-server model.

[2]This can also be motivated through learning over local wireless mesh (or ad hoc) networks.

[3]Acronym stands for Sparsified and Quantized Action Regulated Momentum Stochastic Gradient Descent. See Algorithm 1 for a description of SQuARM-SGD.

ment of stochastic gradients [19]–[21] (presented in Section III-B),(ii) A relatively weaker set of assumptions on the node variance and the gradient dissimilarity across nodes (presented in Section III-A). Specifically, the bounds on the variance and the gradient dissimilarity depend on the local geometry of the true gradients; see Assumption 2 for the bounded variance assumption and Assumption 3 for the bounded gradient dissimilarity assumption. Both these assumptions are strictly weaker than assuming uniform bounds on the respective quantities; see Remark 1 for a detailed discussion. For assumptions set (i), we show a convergence rate of $\mathcal{O}\left(1/\sqrt{nT}\right)$ for smooth convex and non-convex objectives, where n is the number of worker nodes and T is the number of iterations, thus matching the convergence rate of vanilla distributed SGD. Similarly, for the weaker assumption set (ii), we show a convergence rate of $\mathcal{O}\left(1/\sqrt{T}\right)$ for smooth non-convex objectives. We note that compression and event triggered communication do affect our convergence rate expressions for results in both sets of assumptions, but they appear only in the higher order terms; thus, for a large enough T, we can converge at the same rate as that of distributed vanilla SGD while enjoying the savings in communication from our method essentially for free; see Theorem 1 and Theorem 2 and comments after that for details. As mentioned earlier, we use Nesterov's momentum in SQuARM-SGD and theoretically analyze its convergence rate; a first theoretical analysis of convergence of such compressed gradient updates with momentum in the decentralized setting. In order to achieve this, we had to solve several technical difficulties; see Section IV and also the related work below. Our numerical results for decentralized training of ResNet20 [23] model on CIFAR-10 [24] dataset shows that including momentum updates as in SQuARM-SGD can lead to around 2% increase in test accuracy performance in comparison to the recently proposed communication efficient algorithms CHOCO-SGD [20] or SPARQ-SGD [19] which do not use momentum.

Related work. Communication-efficient decentralized training has received recent attention; see [18]–[20], [25]–[30] and references therein. CHOCO-SGD proposed by [20], [21] was the first to perform arbitrary compressed training for decentralized optimization by considering sparsification or quantization of the model parameters. Recently, in [19] we proposed SPARQ-SGD incorporating compression using both sparsification and quantization and also event-driven communication with local iterations to save on communicated bits. We remark that [19]–[21] rely on (a strong) assumption of bounded second moment of stochastic gradients for their theoretical analysis and do not incorporate momentum updates, which has been shown to empirically improve generalization performance in deep learning applications [28], [31]. Our convergence analyses are very different and more involved than CHOCO-SGD or SPARQ-SGD, as we rely on a much weaker set of assumptions and provide our analyses using virtual sequences, specifically, to handle the use of momentum. Use of local iterations in decentralized setting with a weaker set of assumptions similar to ours has been considered recently in [32], however, without any compression of updates, and importantly, without incorporating momentum in the theoretical analysis. The use of local iterations with momentum updates in decentralized setting has been studied in [33], but without any compression of exchanged information and with a stronger set of assumptions. [34] studied momentum SGD with compressed updates (but no local iterations or event-triggering) for the *distributed* setting only, assuming that all workers have access to unbiased gradients. Extending the analysis to the *decentralized* setting (where different workers may have local data, potentially generated from different distributions) while incorporating momentum, compression, local iterations, and event triggered communication[4] (as in SQuARM-SGD) while assuming a weaker set of assumptions than existing works poses several challenges; see Section IV for a detailed discussion. The idea of event-triggering has been explored in the control community [35]–[39] and in the optimization literature [40]–[42]. These papers focus on continuous-time, deterministic optimization algorithms for convex problems; in contrast, our event-driven stochastic gradient descent algorithm is for both convex and general (non-convex) smooth objectives, e.g., neural network training for large-scale deep learning. [43] proposed an adaptive scheme to skip gradient computations in a *distributed* setting for *deterministic* gradients; moreover, their focus is on saving communication rounds, without compressed communication. To the best of our knowledge, ours is the first paper to develop and analyze convergence of momentum-based decentralized stochastic optimization, using compressed lazy communication (as described earlier). Moreover, our numerics demonstrate better test-accuracy performance compared to recently proposed methods for communication efficiency on account of using momentum updates.

Paper organization. The problem setup and our algorithm SQuARM-SGD are described in Section II. Section III provides two sets of convergence results, one with weak assumptions (Theorem 1), and the other (a slightly general result) with strong assumptions (Theorem 2). We prove Theorem 1 in Section V (which is a novel analysis and the main technical contribution of our paper) and defer the proof of Theorem 2 to the supplementary material. Section VI gives numerical results comparing our algorithm to the state-of-the-art. Omitted proofs/details are provided in appendices.

II. Problem Setup and Our Algorithm

We first formalize the decentralized optimization setting that we work with and set up the notation we follow throughout the paper. Consider an undirected connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = [n] := \{1, 2, \ldots, n\}$, where node $i \in [n]$ corresponds to worker i and we denote the neighbors of node i by $\mathcal{N}_i := \{(i,j) : (i,j) \in \mathcal{E}\}$. To each node $i \in [n]$, we associate a dataset $\mathcal{D}_i$ and an objective function $f_i : \mathbb{R}^d \to \mathbb{R}$. We allow the datasets and objective functions to be different for each node and assume that for $i \in [n]$, the objective function

[4]Event-triggered communication with compression and local iterations is also considered in [19], however, with the strong bounded second moment gradient assumption and without momentum updates in the theoretical analysis. Relaxing the assumptions and incorporating momentum significantly changes the convergence analysis (see Section IV).

f_i has the form $f_i(\mathbf{x}) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[F_i(\mathbf{x}, \xi_i)]$ where $\xi_i \sim \mathcal{D}_i$ denotes a random sample from $\mathcal{D}_i$, $\mathbf{x}$ denotes the parameter vector, and $F_i(\mathbf{x}, \xi_i)$ denotes the risk associated with sample ξ_i with respect to (w.r.t.) the parameter vector $\mathbf{x}$. Consider the following empirical risk minimization problem, where $f : \mathbb{R}^d \to \mathbb{R}$ is called the global objective function:

$$\arg\min_{\mathbf{x}\in\mathbb{R}^d} \Big(f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \Big), \tag{1}$$

The nodes in $\mathcal{G}$ wish to minimize (1) collaboratively in a communication-efficient manner while incorporating momentum updates of worker nodes.

We now state the notation relevant to describing our algorithm. Let $\mathbf{W} \in \mathbb{R}^{n\times n}$ denote the connectivity matrix of $\mathcal{G}$, where for every $(i,j) \in \mathcal{E}$, the (i,j)'th entry of $\mathbf{W}$ denotes the weight w_{ij} on the edge (i,j) – e.g., w_{ij} may represent the strength of the connection on the edge (i,j) – and for other pairs $(i,j) \notin \mathcal{E}$, the weight w_{ij} is zero. We assume that $\mathbf{W}$ is symmetric and doubly stochastic, which means it has non-zero entries with each row and column summing up to 1. Consider the ordered eigenvalues of $\mathbf{W}$, $|\lambda_1(\mathbf{W})| \geq |\lambda_2(\mathbf{W})| \geq \ldots \geq |\lambda_n(\mathbf{W})|$. For such a $\mathbf{W}$ associated with a connected graph $\mathcal{G}$, it is known that $\lambda_1(\mathbf{W}) = 1$ and $\lambda_i(\mathbf{W}) \in (-1, 1)$ for all $i \in \{2, \ldots, n\}$. The spectral gap $\delta \in (0, 1]$ is defined as $\delta := 1 - |\lambda_2(\mathbf{W})|$. Simple matrices $\mathbf{W}$ having $\delta \in (0, 1]$ are known to exist for connected graphs [21].

To achieve compression on the communication exchanged between workers, we use arbitrary compression operators as defined next.

Definition 1 (Compression, [5])**.** *A (possibly randomized) function $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is called a* compression *operator, if there exists a positive constant $\omega \in (0, 1]$, such that for every $\mathbf{x} \in \mathbb{R}^d$:*

$$\mathbb{E}_{\mathcal{C}}[\|\mathbf{x} - \mathcal{C}(\mathbf{x})\|_2^2] \leq (1-\omega)\|\mathbf{x}\|_2^2, \tag{2}$$

where expectation is taken over the randomness of $\mathcal{C}$. We assume that $\mathcal{C}(\mathbf{0}) = \mathbf{0}$.

We now list some important sparsifiers and quantizers following the above definition of a compression operator:
(i) Top_k and $Rand_k$ sparsifiers (where only k entries are selected and the rest are set to zero) with $\omega = k/d$ [5], (ii) Stochastic quantizer Q_s from [7][5] with $\omega = (1 - \beta_{d,s})$ for $\beta_{d,s} < 1$, and (iii) Deterministic quantizer $\frac{\|\mathbf{x}\|_1}{d} Sign(\mathbf{x})$ from [10] with $\omega = \frac{\|\mathbf{x}\|_1^2}{d\|\mathbf{x}\|_2^2}$. For $Comp_k \in \{Top_k, Rand_k\}$, the following are compression operators[6]: (iv) $\frac{1}{(1+\beta_{k,s})} Q_s(Comp_k)$ with $\omega = \left(1 - \frac{k}{d(1+\beta_{k,s})}\right)$ for any $\beta_{k,s} \geq 0$, and (v) $\frac{\|Comp_k(\mathbf{x})\|_1 Sign(Comp_k(\mathbf{x}))}{k}$ with $\omega = \max\left\{\frac{1}{d}, \frac{k}{d}\left(\frac{\|Comp_k(\mathbf{x})\|_1^2}{d\|Comp_k(\mathbf{x})\|_2^2}\right)\right\}$ [12].

[5] $Q_s : \mathbb{R}^d \to \mathbb{R}^d$ is a stochastic quantizer, if for every $\mathbf{x} \in \mathbb{R}^d$, we have (i) $\mathbb{E}[Q_s(\mathbf{x})] = \mathbf{x}$ and (ii) $\mathbb{E}[\|\mathbf{x} - Q_s(\mathbf{x})\|_2^2] \leq \beta_{d,s}\|\mathbf{x}\|_2^2$. Q_s from [7] satisfies this definition with $\beta_{d,s} = \min\left\{\frac{d}{s^2}, \frac{\sqrt{d}}{s}\right\}$.

[6] [12] show that the composition of sparsification and quantization operators is also a valid compression operator, outperforming its individual components in terms of communication savings while maintaining similar performance.

Algorithm 1 SQuARM-SGD: Sparsified and Quantized Action Regulated Momentum SGD

Parameters: $G = ([n], E)$, $\mathbf{W}$, Compression operator $\mathcal{C}$

Initialize: For every $i \in [n]$, set arbitrary $\mathbf{x}_i^{(0)} \in \mathbb{R}^d$, $\hat{\mathbf{x}}_i^{(0)} := \mathbf{0}$, $\mathbf{v}_i^{(-1)} := \mathbf{0}$. Fix the momentum coefficient β, consensus step-size γ, learning rate η, triggering thresholds $\{c_t\}_{t=0}^{T}$, and synchronization set $\mathcal{I}_T$.
for $t = 0$ **to** $T-1$ in parallel for all workers $i \in [n]$ **do**
  Sample $\xi_i^{(t)}$, compute stochastic gradient $\mathbf{g}_i^{(t)} := \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})$
  $\mathbf{v}_i^{(t)} = \beta\mathbf{v}_i^{(t-1)} + \mathbf{g}_i^{(t)}$
  $\mathbf{x}_i^{(t+\frac{1}{2})} := \mathbf{x}_i^{(t)} - \eta(\beta\mathbf{v}_i^{(t)} + \mathbf{g}_i^{(t)})$
  if $(t+1) \in I_T$ **then**
    for neighbors $j \in \mathcal{N}_i \cup i$ **do**
      if $\|\mathbf{x}_i^{(t+\frac{1}{2})} - \hat{\mathbf{x}}_i^{(t)}\|_2^2 > c_t\eta^2$ **then**
        Compute $\mathbf{q}_i^{(t)} := \mathcal{C}(\mathbf{x}_i^{(t+\frac{1}{2})} - \hat{\mathbf{x}}_i^{(t)})$
        Send $\mathbf{q}_i^{(t)}$ and receive $\mathbf{q}_j^{(t)}$
      else
        Send $\mathbf{0}$ and receive $\mathbf{q}_j^{(t)}$
      end if
      $\hat{\mathbf{x}}_j^{(t+1)} := \mathbf{q}_j^{(t)} + \hat{\mathbf{x}}_j^{(t)}$
    end for
    $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t+\frac{1}{2})} + \gamma \sum_{j\in\mathcal{N}_i} w_{ij}(\hat{\mathbf{x}}_j^{(t+1)} - \hat{\mathbf{x}}_i^{(t+1)})$
  else
    $\hat{\mathbf{x}}_i^{(t+1)} = \hat{\mathbf{x}}_i^{(t)}$, $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t+\frac{1}{2})}$ for all $i \in [n]$
  end if
end for

A. Our Algorithm: SQuARM-SGD

We propose SQuARM-SGD to minimize (1), which is a decentralized algorithm that combines compression and Nesterov's momentum, together with event-driven communication exchange, where compression is achieved by sparsifying *and* quantizing the exchanges. Each worker is required to complete a fixed number of *local SGD* steps with *momentum*, and communicate *compressed* updates to its neighbors when there is a *significant change* in its local parameters since the last communication round.

To realize exchange of compressed parameters between workers, for each node $i \in [n]$, all nodes $j \in \mathcal{N}_i$ maintain an estimate $\hat{\mathbf{x}}_i$ of $\mathbf{x}_i$, so, each node $i \in [n]$ has access to $\hat{\mathbf{x}}_j$ for all $j \in \mathcal{N}_i$. Our algorithm runs for T iterations and the set of synchronization indices is defined as $\mathcal{I}_T = \{0, H, 2H \ldots, mH, \ldots\} \subseteq [T]$ for some constant $H \in \mathbb{N}$, which are same for all workers and denote the time steps at which workers are allowed to communicate, provided they satisfy a triggering condition.[7]

For a given connected graph $\mathcal{G}$ with connectivity matrix $\mathbf{W}$, we first initialize a consensus step-size γ (see Theorem 1 for definition), momentum factor β, learning rate η, triggering threshold sequence $\{c_t\}_{t=0}^{T}$, and momentum vector $\mathbf{v}_i$ for each node i initialized to $\mathbf{0}$. We initialize the copies of all the nodes $\hat{\mathbf{x}}_i = \mathbf{0}$ and allow each node to communicate in the first round. At each time step t, each worker $i \in [n]$ samples a stochastic gradient $\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i)$ and takes a local SGD step

[7] The Zeno phenomenon [35] does not occur in our setup as we have a discrete sampling period as well as a fixed number of local iterations, giving a lower bound to the event intervals of at least H times the sampling period.

on parameter $\mathbf{x}_i^{(t)}$ using Nesterov's momentum to form an intermediate parameter $\mathbf{x}_i^{(t+1/2)}$ (lines 3-5). If the next iteration corresponds to a synchronization index, i.e., $(t+1) \in \mathcal{I}_T$, then each worker checks the triggering condition (line 8). If satisfied, that worker communicates the compressed change in its copy to all its neighbors $\mathcal{N}_i$ (lines 9-10); otherwise, it does not communicate in that round (denoted by 'Send $\mathbf{0}$' in our algorithm for illustration, line 12). After receiving the compressed updates of copies from all its neighbors, the node i updates the locally available copies and its own copy (line 14). With these updated copies, the worker nodes finally take a consensus (line 16) with appropriate weighting decided by entries of $\mathbf{W}$. In the case when $(t+1) \notin \mathcal{I}_T$, the nodes maintain their copies and move on to next iteration (line 18); thus no communication takes place.

Difference from SPARQ-SGD [19]: There are two major differences between this work and our previous work [19] which uses a similar framework of local iterations, compression and triggering to save on communication. Firstly, and most importantly, the results presented in this work do not use any strong assumptions like the bounded second moment of stochastic gradients used in [19]–[21]: Both the variance bound on stochastic gradients as well as the data heterogeneity bound depend on local geometry of the true gradients (and we allow these to scale with the true gradient norm); and thus, neither of them are assumed to be uniformly bounded, as in [19]–[21]. The assumptions in this work are thus much weaker than the ones in existing decentralized literature; see Section IV for details. Working with these relaxed assumptions calls for completely different and much more nuanced analyses to establish the convergence rates as compared to [19]. Secondly, the addition of lines 4-5 in Algorithm 1 which now incorporate momentum calls for a significantly different analysis than [19] to arrive at the convergence rate even if we consider the same set of assumptions. Even though momentum updates are almost always used in practice, incorporating them in convergence analyses in modern large-scale settings with communication constraints has received attention only recently, e.g., for distributed training with compressed update exchanges [34] and for decentralized training without compression or local SGD in [28]. To the best of our knowledge, our work provides the first convergence analysis for compressed decentralized training with momentum using a weaker set of assumptions than existing literature while incorporating the local SGD and event triggered communication framework of [19]. We note the technical challenges that arise and provide a detailed comparison to SPARQ-SGD [19] and other recent works analyzing momentum in Section IV. Furthermore, our experimental results in Section VI show that incorporating momentum can empirically improve the generalization performance of the trained model by about 2-3% when compared to training without momentum.

Memory-efficient version of Algorithm 1: At the first glance, it may seem that in Algorithm 1, every node has to store estimates of all its neighbors' parameters in order to perform the consensus step, which may be impractical in large-scale learning. Note that in the consensus step (line 16), nodes only require the weighted sum of their neighbors' parameters. So, it suffices for each node to store only the weighted sum of all its neighbors' parameters (in addition to its own local parameters and its estimate), and thus avoiding the need to store all neighbor parameters. A memory-efficient version of SQuARM-SGD is given in Appendix I.

Equivalence to error-feedback mechanisms: In Algorithm 1, though nodes do not explicitly perform local error-compensation ([10], [12]), the error-compensation happens implicitly. To see this, note that nodes maintain copies of their neighbors' parameters and update them as $\hat{\mathbf{x}}_j^{(t+1)} = \hat{\mathbf{x}}_j^{(t)} + \mathcal{C}(\mathbf{x}_j^{(t+\frac{1}{2})} - \hat{\mathbf{x}}_j^{(t)})$ (line 14) and then perform consensus (line 16). Thus, the error gets accumulated into $\hat{\mathbf{x}}_j^{(t)}$ and is compensated by the term $\mathcal{C}(\mathbf{x}_j^{(t+\frac{1}{2})} - \hat{\mathbf{x}}_j^{(t)})$ in the next round.

III. Main Results

In this section we provide the convergence results for SQuARM-SGD (Algorithm 1) under two sets of assumptions: We present our results with the weakest set of assumptions available in existing literature in Section III-A and slightly more general results with stronger assumptions in Section III-B.

A. Theoretical Results with Relaxed Assumptions

Assumption 1 (Smoothness)**.** *We assume that each local function f_i for $i \in [n]$ is L-smooth, i.e., $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have $f_i(\mathbf{y}) \leq f_i(\mathbf{x}) + \langle \nabla f_i(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2$.*

Assumption 2 (Bounded Variance)**.** *We assume that there exists finite constants $\sigma, M \geq 0$, such that for all $\mathbf{x} \in \mathbb{R}^d$ we have:*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\xi_i}\|\nabla F_i(\mathbf{x}_i, \xi_i) - \nabla f_i(\mathbf{x}_i)\|_2^2 \leq \sigma^2 + \frac{M^2}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}_i)\|_2^2, \tag{3}$$

where $\nabla F_i(\mathbf{x}, \xi_i)$, $i \in [n]$, denotes an unbiased stochastic gradient, i.e., $\mathbb{E}_{\xi_i}[\nabla F_i(\mathbf{x}, \xi_i)] = \nabla f_i(\mathbf{x})$.

Assumption 3 (Bounded Gradient Dissimilarity)**.** *We assume that there exists finite constants $G \geq 0$ and $B \geq 1$, such that for all $\mathbf{x} \in \mathbb{R}^d$ we have:*

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x})\|_2^2 \leq G^2 + B^2\|\nabla f(\mathbf{x})\|_2^2. \tag{4}$$

These assumptions have appeared in literature before in [32] to study decentralized optimization with local iterations; and we extend their results and analyses by incorporating compression and momentum. This extension posed many fundamental technical difficulties, which we describe in detail in Section IV.

Remark 1 (Comparison with Existing Assumptions)**.** *Assumptions 2, 3 are weaker than assuming uniform bounds on the variance and the gradient dissimilarity:* (i) *The uniform bound on the variance [28], i.e., $\mathbb{E}_{\xi_i}\|\nabla F_i(\mathbf{x}_i, \xi_i) - \nabla f_i(\mathbf{x}_i)\|_2^2 \leq \sigma_i^2$ for all $i \in [n]$, implies Assumption 2 with $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$ and $M = 0$; and* (ii) *The uniform bound on the gradient similarity [28], i.e., $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \leq \kappa^2$,*

implies Assumption 3 with $G = \kappa$ and $B = 1$ – this follows from the identity $\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x})\|_2^2 - \|\nabla f(\mathbf{x})\|_2^2$. Both Assumptions 2 and 3 are weaker than the uniformly bounded second moment assumption $\mathbb{E}_{\xi_i}\|\nabla F_i(\mathbf{x}_i,\xi_i)\|_2^2 \leq G^2$, which has been standard in the stochastic optimization with compressed gradients [5], [12], [20], [34].

Our convergence result (stated below) is for general smooth (non-convex) objectives; and can be readily extended to convex objectives. We derive this result for SQuARM-SGD under Assumptions 1-3 without event-triggered communication; in other words, our analysis is for compressed decentralized momentum SGD with local iterations. We would like to emphasize that incorporating event-triggering component into our analysis can only complicate the calculations and can be done. In order to bring out the novelty of our convergence analysis without adding unnecessary technicality, we present the result in this subsection and its subsequent analysis without incorporating event-triggered communication.

Theorem 1. *Let $\mathcal{C}$ be a compression operator with parameter $\omega \in (0,1]$ and $gap(\mathcal{I}_T) = H$. Consider running SQuARM-SGD for T iterations with consensus step-size $\gamma = \frac{2\delta\omega^3}{4\delta^2\omega^2+\delta^2+128\lambda^2+24\omega^2\lambda^2}$, (where $\lambda = max_i\{1-\lambda_i(\mathbf{W})\}$), momentum coefficient $\beta \in [0,1)$, and constant learning rate $\eta = (1-\beta)\sqrt{\frac{n}{T}}$. Let the algorithm generate $\{\mathbf{x}_i^{(t)}\}_{t=0}^{T-1}$ for $i \in [n]$. Running the algorithm for $T \geq U_0$ for some constant U_0 defined in Appendix C-F, the averaged iterates $\overline{\mathbf{x}}^{(t)} := \frac{1}{n}\sum_{i=0}^{n}\mathbf{x}_i^{(t)}$ satisfy:*

$$\frac{\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(t)})\|_2^2}{T} = \mathcal{O}\left(\frac{J^2+\sigma^2+(M^2+n)G^2}{\sqrt{nT}}\right) + \mathcal{O}\left(\frac{(1-\beta)^2 nH^2((M^2+1)G+\sigma^2)}{T\delta^2\omega^3}\right),$$

where $J^2 < \infty$ is such that $\mathbb{E}[f(\overline{\mathbf{x}}^{(0)})] - f^* \leq J^2$.

We prove Theorem 1 in Section V. Note that we have used simplified convergence rate expressions in the above result, and derive precise rate expressions in Section V.

B. Theoretical Results with Bounded Second Moment of Stochastic Gradients

In this section, we consider a stronger set of assumptions than the ones before along with the smoothness of objectives: (i) *Uniformly bounded variance:* For every $i \in [n]$, we have $\mathbb{E}_{\xi_i}\|\nabla F_i(\mathbf{x},\xi_i) - \nabla f_i(\mathbf{x})\|^2 \leq \sigma_i^2$, for some finite σ_i, where $\nabla F_i(\mathbf{x},\xi_i)$ denotes an unbiased stochastic gradient at worker i with $\mathbb{E}_{\xi_i}[\nabla F_i(\mathbf{x},\xi_i)] = \nabla f_i(\mathbf{x})$. We define $\bar{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$. (ii) *Uniformly bounded second moment:* For every $i \in [n]$, we have $\mathbb{E}_{\xi_i}\|\nabla F_i(\mathbf{x},\xi_i)\|^2 \leq G^2 < \infty$.

Theorem 2. *Let $\mathcal{C}$ be a compression operator with parameter $\omega \in (0,1]$ and $gap(\mathcal{I}_T) = H$. Consider running SQuARM-SGD for T iterations with consensus step-size $\gamma = \frac{2\delta\omega}{64\delta+\delta^2+16\lambda^2+8\delta\lambda^2-16\delta\omega}$, (where $\lambda = max_i\{1-\lambda_i(\mathbf{W})\}$), a threshold sequence $c_t \leq \frac{c_0}{\eta^{1-\epsilon}}$ for all t where $\epsilon \in (0,1)$ and c_0 is a constant, momentum coefficient $\beta \in [0,1)$, and constant learning rate $\eta = (1-\beta)\sqrt{\frac{n}{T}}$. Let the algorithm generate $\{\mathbf{x}_i^{(t)}\}_{t=0}^{T-1}$ for $i \in [n]$. Then, we have:*

- **[Non-convex:]** *For $T \geq \max\{16L^2 n, \frac{8L^2\beta^4 n}{(1-\beta)^2}\}$, the averaged iterates $\overline{\mathbf{x}}^{(t)} := \frac{1}{n}\sum_{i=0}^{n}\mathbf{x}_i^{(t)}$ satisfy:*

$$\frac{\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(t)})\|_2^2}{T} = \mathcal{O}\left(\frac{J^2+\bar{\sigma}^2}{\sqrt{nT}}\right) + \mathcal{O}\left(\frac{c_0 n^{(1+\epsilon)/2}}{\delta^2 T^{(1+\epsilon)/2}} + \frac{nH^2G^2}{T\delta^4\omega^2} + \frac{\beta^4\bar{\sigma}^2}{T(1-\beta)^2}\right),$$

  where $J^2 < \infty$ is such that $\mathbb{E}[f(\overline{\mathbf{x}}^{(0)})] - f^* \leq J^2$.
- **[Convex:]** *If $\{f_i\}_{i\in[n]}$ are convex, then for $T \geq \max\{(8L)^2 n, \frac{(8\beta^2 L)^4 n}{(1-\beta)^2}\}$, we have:*

$$\mathbb{E}[f(\overline{\mathbf{x}}_{avg}^{(T)})] - f^* = \mathcal{O}\left(\frac{\|\overline{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2 + \bar{\sigma}^2}{\sqrt{nT}}\right) + \mathcal{O}\left(\frac{c_0 n^{(1+\epsilon)/2}}{\delta^2 T^{(1+\epsilon)/2}} + \frac{n^{3/4}\beta^2 G^2}{(1-\beta)^{3/2}T^{3/4}} + \frac{nH^2G^2}{\delta^4\omega^2 T}\right),$$

  where $\overline{\mathbf{x}}_{avg}^{(T)} := \frac{1}{T}\sum_{t=0}^{T-1}\overline{\mathbf{x}}^{(t)}$ for $\overline{\mathbf{x}}^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^{(t)}$ and $\mathbf{x}^*$ is an optimizer of f attaining optimal value f^*.

We have used simplified convergence rate expressions in the above results, and provide precise rate expressions in the proofs provided in Appendix E and Appendix F for non-convex and convex objectives, respectively.

C. Effects of parameters on convergence

The factors arising due to communication efficiency – H (and c_0 for Theorem 2) for the event-triggered communication, ω for compression, and δ for the connectivity of the underlying graph – do not affect the dominant terms in convergence rate for either Theorem 1 or Theorem 2 and appear only in the higher order terms. This implies that if we run SQuARM-SGD for sufficiently long, precisely, for at least $T_{w_0} = C_{w_0} \times \left(\frac{n^3}{\delta^4\omega^4}\frac{(1-\beta^2)^2H^4\left[(M^2+1)G+\sigma^2\right]^2}{\left[J^2+\sigma^2+(M^2+n)G^2\right]^2}\right)$ where G, σ, M are defined in the weaker set of assumptions provided in Subsection III-A and C_{w_0} is a sufficiently large constant, then SQuARM-SGD converges at a rate $\mathcal{O}\left(1/\sqrt{T}\right)$. Similarly, if we consider the stronger set of assumptions stated in Subsection III-B, and run SQuARM-SGD for at least $T_{s_0} := C_{s_0} \times \max\left\{\left(\frac{c_0^2 n^{(2+\epsilon)}}{(J^2+\bar{\sigma}^2)^2\delta^4}\right)^{1/\epsilon}, \frac{n}{(J^2+\bar{\sigma}^2)^2}\left(\frac{nG^2H^2}{\omega^2\delta^4} + \frac{\beta^4\bar{\sigma}^2}{(1-\beta)^2}\right)^2\right\}$ iterations for non-convex objectives and for $T_{s_1} := C_{s_1} \times \max\left\{\left(\frac{c_0^2 n^{2+\epsilon}}{\delta^4(\|\overline{\mathbf{x}}^{(0)}-\mathbf{x}^*\|^2+\bar{\sigma}^2)^2}\right)^{1/\epsilon}, \frac{n^3H^4G^2}{\delta^8\omega^4(\|\overline{\mathbf{x}}^{(0)}-\mathbf{x}^*\|^2+\bar{\sigma}^2)^2}, \frac{n^5G^8\beta^8}{(1-\beta)^6(\|\overline{\mathbf{x}}^{(0)}-\mathbf{x}^*\|^2+\bar{\sigma}^2)^4}\right\}$ for convex objectives with sufficiently large constants C_{s_0} and C_{s_1}, respectively, then SQuARM-SGD converges at a rate of $\mathcal{O}\left(1/\sqrt{nT}\right)$. Note that this is the convergence rate of distributed *vanilla* SGD with the same speed-up w.r.t. the number of nodes n in both these settings. Thus, we essentially converge at the same rate as that of vanilla SGD, while saving significantly in terms of total communicated bits; this can also be seen in our numerical results in Section VI.

IV. Preliminaries

In this section, we first establish a matrix notation which would be used throughout the proofs. We then state SQuARM-SGD in matrix notation (which is equivalent to Algorithm 1) and list important facts regarding our updates. We conclude this section with a brief discussion of technical challenges involved in the proofs.

Matrix notation. Consider the set of parameters $\{\mathbf{x}_i^{(t)}\}_{i=1}^n$ at all nodes at timestep t as well as the estimates of the parameters $\{\hat{\mathbf{x}}_i^{(t)}\}_{i=1}^n$. The matrix notation is given by:

$$\mathbf{X}^{(t)} := [\mathbf{x}_1^{(t)}, \dots, \mathbf{x}_n^{(t)}] \in \mathbb{R}^{d\times n}$$
$$\hat{\mathbf{X}}^{(t)} := [\hat{\mathbf{x}}_1^{(t)}, \dots, \hat{\mathbf{x}}_n^{(t)}] \in \mathbb{R}^{d\times n}$$
$$\bar{\mathbf{X}}^{(t)} := [\bar{\mathbf{x}}^{(t)}, \dots, \bar{\mathbf{x}}^{(t)}] \in \mathbb{R}^{d\times n}$$
$$\mathbf{V}^{(t)} := [\mathbf{v}_1^{(t)}, \mathbf{v}_2^{(t)}, \dots, \mathbf{v}_n^{(t)}] \in \mathbb{R}^{d\times n}$$
$$\boldsymbol{\nabla F}(\mathbf{X}^{(t)}, \boldsymbol{\xi}^{(t)}) := [\nabla F_1(\mathbf{x}_1^{(t)}, \xi_1^{(t)}), \dots, \nabla F_n(\mathbf{x}_n^{(t)}, \xi_n^{(t)})] \in \mathbb{R}^{d\times n}$$

Here, $\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})$ denotes the stochastic gradient at node i at timestep t and the vector $\bar{\mathbf{x}}^{(t)} := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^{(t)}$ denotes the average of node parameters at time t. Let $\Gamma^{(t)} \subseteq [n]$ be the set of nodes that do not communicate at time t. We define $\mathbf{P}^{(t)} \in \mathbb{R}^{n\times n}$, a diagonal matrix with $\mathbf{P}_{ii}^{(t)} = 0$ for $i \in \Gamma^{(t)}$ and $\mathbf{P}_{ii}^{(t)} = 1$ otherwise.

SQuARM-SGD in matrix notation. Consider Algorithm 1 with synchronization indices given by the set $\mathcal{I}_T = \{0, H, 2H \dots, mH, \dots\} \subseteq [T]$ for some constant $H \in \mathbb{N}$. Using the above notation, the sequence of parameters' updates from synchronization index mH to $(m+1)H$ is:

$$\mathbf{V}^{(t)} = \beta\mathbf{V}^{(t-1)} + \boldsymbol{\nabla F}(\mathbf{X}^{(t)}, \boldsymbol{\xi}^{(t)}) \tag{5}$$

$$\mathbf{X}^{((m+1/2)H)} = \mathbf{X}^{I_{(t)}} - \sum_{t'=mH}^{(m+1)H-1} \eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})) \tag{6}$$

$$\hat{\mathbf{X}}^{((m+1)H)} = \hat{\mathbf{X}}^{(mH)} + \mathcal{C}((\mathbf{X}^{((m+1/2)H)} - \hat{\mathbf{X}}^{(mH)})\mathbf{P}^{((m+1)H-1)}) \tag{7}$$

$$\mathbf{X}^{((m+1)H)} = \mathbf{X}^{((m+1/2)H)} + \gamma\hat{\mathbf{X}}^{((m+1)}(\mathbf{W} - \mathbf{I}) \tag{8}$$

where $\mathcal{C}(.)$ denotes the compression operator applied column-wise to the argument matrix and $\mathbf{I}$ is the identity matrix. Note that in the update rule for $\hat{\mathbf{X}}^{((m+1)H)}$, we used (i) the fact that $\mathbf{P}$ is a diagonal matrix and that $\mathcal{C}$ is applied column-wise to write $\mathcal{C}(\mathbf{X}^{((m+1/2)H)} - \hat{\mathbf{X}}^{(mH)})\mathbf{P}^{((m+1)H-1)} = \mathcal{C}((\mathbf{X}^{((m+1/2)H)} - \hat{\mathbf{X}}^{(mH)})\mathbf{P}^{((m+1)H-1)})$, and (ii) that $\hat{\mathbf{X}}^{((m+1)H-1)} = \hat{\mathbf{X}}^{(mH)}$, because $\hat{\mathbf{X}}$ does not change in between the synchronization indices.

We now note some useful properties of the iterates in matrix notation which would be used throughout the paper:

1) Since $\mathbf{W} \in [0,1]^{n\times n}$ is a doubly stochastic matrix, we have: $\mathbf{W} = \mathbf{W}^T, \mathbf{W}\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^T\mathbf{W} = \mathbf{1}^T$ (where $\mathbf{1}$ is the all ones vector in $\mathbb{R}^n$). This also gives us:
$$\bar{\mathbf{X}}^{(t)} := \mathbf{X}^{(t)}\frac{1}{n}\mathbf{1}\mathbf{1}^T, \qquad \bar{\mathbf{X}}^{(t)}\mathbf{W} = \bar{\mathbf{X}}^{(t)} \tag{9}$$
where the first expression follows from the definition of $\bar{\mathbf{X}}^{(t)}$ and the second expression follows because $\mathbf{W}\frac{\mathbf{1}\mathbf{1}^T}{n} = \frac{\mathbf{1}\mathbf{1}^T}{n}\mathbf{W} = \frac{1}{n}\mathbf{1}\mathbf{1}^T$.

2) The average of the iterates in Algorithm 1 follows :
$$\bar{\mathbf{X}}^{(t+1)} = \bar{\mathbf{X}}^{(t+\frac{1}{2})} + \mathbf{1}_{(t+1)\in\mathcal{I}_T}\left[\gamma\hat{\mathbf{X}}^{(t+1)}(\mathbf{W}-\mathbf{I})\frac{1}{n}\mathbf{1}\mathbf{1}^T\right] = \bar{\mathbf{X}}^{(t+\frac{1}{2})} \tag{10}$$
where $\mathcal{I}_T$ denotes the set of synchronization indices of Algorithm 1. We use $(\mathbf{W}-\mathbf{I})\frac{1}{n}\mathbf{1}\mathbf{1}^T = \mathbf{W}\frac{\mathbf{1}\mathbf{1}^T}{n} - \frac{\mathbf{1}\mathbf{1}^T}{n} = \mathbf{0}$.

Proposition 1 (Variance Reduction with Independent Samples)**.** *Consider the variance bound* (3) *on the stochastic gradient for nodes. If $\boldsymbol{\xi}^{(t)} = \{\xi_1^{(t)}, \xi_2^{(t)}, \dots, \xi_n^{(t)}\}$ denotes the collection of independent stochastic samples for the nodes at any time-step t. Then we have:*

$$\mathbb{E}_{\boldsymbol{\xi}^{(t)}}\left\|\frac{1}{n}\sum_{i=1}^n \nabla\big(F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})\big)\right\|^2 \le \frac{\sigma^2}{n} + \frac{M^2}{n^2}\sum_{i=1}^n \left\|\nabla f_i(\mathbf{x}_i^{(t)})\right\|_2^2. \tag{11}$$

Proposition 2. *For any t, $\mathbb{E}\left\|\mathbf{V}^{(t)}\right\|_F^2$ is bounded as follows:*

$$(1-\beta)\mathbb{E}\|\mathbf{V}^{(t)}\|_F^2 \le \Lambda^{(t)} := \sum_{k=0}^t \beta^{t-k}\mathbb{E}\|\boldsymbol{\nabla F}(\mathbf{X}^{(k)}, \boldsymbol{\xi}^{(k)})\|_F^2 \tag{12}$$

We prove the above propositions in Appendix B.

Technical Challenges: We focus on two major aspects of our work to compare with existing literature: (i) Analysis of compressed decentralized training with triggered communication with mild assumptions. (ii) Performing the resulting analysis by taking into account the momentum updates.

The assumption on bounded second moment of stochastic gradients is commonly used in communication efficient decentralized training literature [19]–[22], and is also used to derive the result of Theorem 2 in our paper. However, this assumption can be quite strong for settings where the data distribution among clients is heterogeneous, as the gradient dissimilarity between clients can be bounded trivially using the second moment bound (see the note on comparison of assumptions in Remark 1 on page 4). In contrast, in Theorem 1, we work with a much weaker set of assumptions (see Section III-A) by not assuming any uniform bound on norm of stochastic gradients, and further allow both the gradient diversity and the variance of stochastic gradients to scale with the norm of gradients compared to existing works [28]. Performing the analyses with these relaxed assumptions is challenging, as it requires us to carefully consider the error due to quantization and local iterations per communication round and construct a recursion equation for it (see Lemmas 2, 3 on page 8) and then delicately handle the recursion to bound the error for any time index (see Lemma 4 on page 8). We remark that the assumptions considered for Theorem 1 in our paper have appeared in literature before in [32] to study decentralized optimization with only local iterations; our work is a significant extension of their results and analyses as we incorporate compression and momentum while achieving a convergence rate of $\mathcal{O}\left(1/\sqrt{T}\right)$.

While momentum updates are almost always used in practice to empirically speedup the training process and to improve generalization performance, it has remained unclear whether convergence with linear speedup with number of nodes n (as in the case of SGD without momentum [12], [19], [32], [44]) is still possible when using momentum. Recently, [28], [34] provided a positive answer to this question, where [28] studies local SGD with momentum in a decentralized setup, but *without* any compressed or event-triggered communication, and [34] studies compressed *distributed* SGD with momentum for non-convex objectives, but without local iterations or event-triggered communication. Our result in Theorem 2 is the first to provide convergence rates showing linear speedup with n for compressed *decentralized* optimization using momentum while incorporating local iteration and triggered communication in the analysis (see Section III-B for the convergence result and the assumptions made). To achieve this, our convergence proofs require the use of virtual sequences as defined in (13) on page 7. Proving convergence results using virtual sequences has been promising lately in stochastic optimization; see, for example, [5], [6], [10], [12], [28], [34].

We would like to emphasize that even without momentum and local iterations, analyzing compression in decentralized optimization [19]–[21] (whose analysis does not require virtual sequences) is significantly more involved and requires different technical tools than analyzing compression in distributed optimization [6], [10]. One of the main reasons for this is as follows: In a decentralized setup, we need to separately show that nodes eventually reach to the same parameters (i.e., consensus happens), which happens trivially in a distributed setup, because in each iteration all worker nodes have the same parameters sent by the master node. On top of that, incorporating momentum updates (which has only been analyzed with compression in distributed setups so far) in decentralized setting is non-trivial and gives similar challenges.

As a consequence, it is not surprising that our proofs are fundamentally different and significantly more challenging from existing works, including [19]–[21], [28], [32], [34], as we study momentum updates for decentralized setup with compression, local iterations and event-triggered communication to save on communication bits. Unlike [34], we allow *heterogeneous* setting, where different nodes may have different datasets. Moreover, with all these, we achieve vanilla SGD like convergence rates for non-convex and convex objectives.

V. Results with Relaxed Assumptions: Proof of Theorem 1

In order to prove Theorem 1, we define a virtual sequence $\widetilde{\mathbf{x}}_i^{(t)}$ for each node $i \in [n]$, as follows:

$$\widetilde{\mathbf{x}}_i^{(t)} = \mathbf{x}_i^{(t)} - \frac{\eta\beta^2}{(1-\beta)}\mathbf{v}_i^{(t-1)}; \qquad \widetilde{\mathbf{x}}_i^{(0)} := \mathbf{x}_i^{(0)}. \tag{13}$$

This remaining section is divided into seven subsections. In Section V-A, we derive an SGD like update rule for the virtual sequence. In Section V-B, we provide a proof-outline of Theorem 1. The remaining subsections are dedicated to prove the lemmas stated in the proof outline given in Section V-B.

A. Deriving an SGD-Like Update Rule for the Virtual Sequene

In (13), $\mathbf{x}_i^{(t)}$ is the true local parameter at node i at the t'th iteration, which is equal to (see line 16 of Algorithm 1):

$$\mathbf{x}_i^{(t)} = \mathbf{x}_i^{(t-\frac{1}{2})} + \mathbb{1}_{\{t\in\mathcal{I}_T\}}\gamma\sum_{j=1}^n w_{ij}(\hat{\mathbf{x}}_j^{(t)} - \hat{\mathbf{x}}_i^{(t)}),$$

where $\mathbf{x}_i^{(t-\frac{1}{2})} = \mathbf{x}_i^{(t-1)} - \eta(\beta\mathbf{v}_i^{(t-1)} + \nabla F_i(\mathbf{x}_i^{(t-1)}, \xi_i^{(t-1)}))$ (line 5 in Algorithm 1). Note that we changed the summation from $j \in \mathcal{N}_i$ to $j = 1$ to n; this is because $w_{ij} = 0$ whenever $j \notin \mathcal{N}_i$.

Let $\overline{\mathbf{x}}^{(t)} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^{(t)}$ denote the average of the local iterates at time t. Now we argue that $\overline{\mathbf{x}}^{(t)} = \overline{\mathbf{x}}^{(t-\frac{1}{2})}$. This trivially holds when $t \notin \mathcal{I}_T$. For the other case, i.e., $t \in \mathcal{I}_T$, this follows because $\sum_{i=1}^n\sum_{j=1}^n w_{ij}(\hat{\mathbf{x}}_j^{(t)} - \hat{\mathbf{x}}_i^{(t)}) = 0$, which uses the fact that W is a doubly stochastic matrix. Thus, we have

$$\overline{\mathbf{x}}^{(t)} = \overline{\mathbf{x}}^{(t-1)} - \frac{\eta}{n}\sum_{i=1}^n\left(\beta\mathbf{v}_i^{(t-1)} + \nabla F_i(\mathbf{x}_i^{(t-1)}, \xi_i^{(t-1)})\right). \tag{14}$$

Taking average over all the nodes in (13) and defining $\widetilde{\mathbf{x}}^{(t)} := \frac{1}{n}\sum_{i=1}^n \widetilde{\mathbf{x}}_i^{(t)}$, we get

$$\widetilde{\mathbf{x}}^{(t)} = \overline{\mathbf{x}}^{(t)} - \frac{\eta\beta^2}{(1-\beta)}\frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^{(t-1)}.$$

We now note a recurrence relation for the sequence $\widetilde{\mathbf{x}}^{(t+1)}$:

$$\begin{aligned}
\widetilde{\mathbf{x}}^{(t+1)} &= \overline{\mathbf{x}}^{(t+1)} - \frac{\eta\beta^2}{(1-\beta)}\frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^{(t)} \\
&= \overline{\mathbf{x}}^{(t)} - \frac{\eta}{n}\sum_{i=1}^n\left(\beta\mathbf{v}_i^{(t)} + \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right) - \frac{\eta\beta^2}{(1-\beta)}\frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^{(t)} \\
&= \overline{\mathbf{x}}^{(t)} - \frac{\eta}{n}\sum_{i=1}^n \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) - \left(\eta\beta + \frac{\eta\beta^2}{(1-\beta)}\right)\frac{1}{n}\sum_{i=1}^n \mathbf{v}_i^{(t)} \\
&= \overline{\mathbf{x}}^{(t)} - \frac{\eta}{n}\sum_{i=1}^n \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) - \frac{\eta\beta}{(1-\beta)}\frac{1}{n}\sum_{i=1}^n \beta\mathbf{v}_i^{(t-1)} \\
&\quad - \frac{\eta\beta}{(1-\beta)}\frac{1}{n}\sum_{i=1}^n \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) \\
&= \widetilde{\mathbf{x}}^{(t)} - \frac{\eta}{(1-\beta)}\frac{1}{n}\sum_{i=1}^n \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})
\end{aligned} \tag{15}$$

B. Proof Outline of Theorem 1

The proof is divided into four lemmas. The first lemma (stated in Lemma 1) derives the required convergence bound, however, the RHS depends on the deviation of local parameter vectors from the average parameter vector (i.e., $\Xi^{(t)} := \sum_{i=1}^n \mathbb{E}\left\|\mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)}\right\|_2^2$), which we have to bound. The remaining three lemmas are dedicated to bounding this quantity.

Note that bounding this in the *distributed* setup is not difficult, as at synchronization indices all parameters are the same because it is coordinated by a central server. This means that at any time index $t \in [T]$, there is always a time index $t - H \le t' \le t$ when $\mathbf{x}_i^{(t')}$ for all $i \in [n]$ are the same, and we

have a reference point no too far in the past. However, in the decentralized setup, there is no central server for coordinating the updates, and hence there is no reference point in the past when the local parameters are the same. Moreover, our assumptions are arguably the weakest in literature, and we also are working with compression and momentum updates. Thus, bounding $\Xi^{(t)}$ in our setup is highly non-trivial, and is one of the major technical contributions of our work.

Lemma 1. *Under the setting of Theorem 1, when* $\eta \leq \min\left\{\frac{2(1-\beta)^3}{9\beta^4}, \frac{2(1-\beta)^2}{3\beta^2 L}\sqrt{\frac{n}{M^2+n}}, \frac{(1-\beta)^2}{6\beta^2 LB}\sqrt{\frac{n}{2(M^2+n)}}\right\}$*, we get:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|_2^2 \leq \frac{16\eta L}{(1-\beta)}\Big(\frac{\sigma^2+2(M^2+n)G^2}{n}\Big) + \frac{16(1-\beta)(f(\overline{\mathbf{x}}^{(0)})-f^*)}{\eta T} + \frac{64L^2}{n}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_i^{(t)}-\overline{\mathbf{x}}^{(t)}\|_2^2$$

We provide a proof for Lemma 1 in Section V-C.

Consider any arbitrary $t \in [T]$. We bound $\Xi^{(t)} = \sum_{i=1}^n \mathbb{E}\left\|\mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)}\right\|_2^2$ via another quantity $S^{(t)}$ defined as $S^{(t)} := \Xi^{(t)} + \mathbb{E}\|\mathbf{X}^{(t)} - \widehat{\mathbf{X}}^{((m+1)H)}\|_F^2$, where $m = \lfloor \frac{t}{H} \rfloor - 1$. We derive two upper bounds on $S^{(t)}$ depending on the value of t. Note that in both the following lemmas, $m = \lfloor \frac{t}{H} \rfloor - 1$.

Lemma 2. *Consider any* $t \in [T]$*. Then for* $m = \lfloor \frac{t}{H} \rfloor - 1$*, we have the following bound for* $(m+1)H \leq t \leq (m+2)H-1$*:*

$$S^{(t)} \leq \left(1-\frac{\gamma\delta}{4}\right)S^{(mH)} + 2c_1\eta^2H^2n\left(2(M^2+1)G^2+\sigma^2\right) + c_1\eta^2H\beta^2\sum_{t'=mH}^{t-1}\mathbb{E}\|\mathbf{V}^{(t')}\|_F^2 + 2c_1\eta^2H(M^2+1)L^2\sum_{t'=mH}^{t-1}S^{(t')} + 2c_1\eta^2H(M^2+1)nB^2\sum_{t'=mH}^{t-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|_2^2,$$

where $c_1 \leq 2(1+\frac{\gamma\delta}{4})\left(\frac{3}{\gamma\delta}+\frac{9\lambda^2}{\delta^2}+\frac{45\gamma\lambda^2}{\delta\omega}+\frac{104\gamma^2\lambda^2}{\omega^2}+\frac{4}{\omega}-2\right)+4(1+\frac{4}{\gamma\delta})$.

We provide a proof of Lemma 2 in Section V-E.

Lemma 3. *For* $mH \leq \hat{t} < (m+1)H$*, we have:*

$$S^{(\hat{t})} \leq \left(1+\frac{\gamma\delta}{4}\right)S^{(mH)} + 2c_1\eta^2H^2n\left(2(M^2+1)G^2+\sigma^2\right) + c_1\eta^2H\beta^2\sum_{t'=mH}^{\hat{t}-1}\mathbb{E}\|\mathbf{V}^{(t')}\|_F^2 + 2c_1\eta^2H(M^2+1)L^2\sum_{t'=mH}^{\hat{t}-1}S^{(t')} + 2c_1\eta^2H(M^2+1)nB^2\sum_{t'=mH}^{\hat{t}-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|_2^2,$$

where c_1 *is exactly the same as in Lemma 2.*

We prove Lemma 3 in Section V-F. Using both these lemmas, we will be able to bound $\Xi^{(t)}$. We state the result in the following lemma, which we prove in Section V-G.

Lemma 4. *Under setting of Theorem 1, when* $\eta \leq \min\left\{\sqrt{\frac{\gamma\delta}{512c_1H^2(M^2+1)L^2}}, \sqrt{\frac{\alpha(1-\beta)}{128DH(M^2+1)L^2}}\right\}$*, we have:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}_i^{(t)}-\overline{\mathbf{x}}^{(t)}\right\|_2^2 = \frac{1}{T}\sum_{t=0}^{T-1}S^{(t)} \leq 2\eta^2J_1 + 2\eta^2J_2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|^2,$$

where $J_1 = \left(\frac{32HA}{\alpha}+\left(\frac{32DH}{\alpha}\right)\left(\frac{2(M^2+1)nG^2+n\sigma^2}{(1-\beta)}\right)\right)$ *and* $J_2 = \left(\frac{32CH}{\alpha}+\left(\frac{32DH}{\alpha}\right)\frac{2(M^2+1)nB^2}{(1-\beta)}\right)$*, where* $A = 2c_1H^2n\left(2(M^2+1)G^2+\sigma^2\right)$*,* $C = 2c_1H(M^2+1)nB^2$*, and* $D = \frac{c_1H\beta^2}{(1-\beta)}$*, and* c_1 *is exactly the same as in Lemma 2.*

Substituting the bounds from Lemma 4 into Lemma 1 and choosing $\eta = (1-\beta)\sqrt{\frac{n}{T}}$ (and running the algorithm for a sufficiently long time) completes the proof. Details with exact numbers are provided in Appendix C-F.

C. Proof of Lemma 1

Consider the quantity $\mathbb{E}_{\xi_{(t)}}[f(\widetilde{\mathbf{x}}^{(t+1)})]$ where expectation is taken w.r.t. the sampling at time t. From the recurrence relation of the virtual sequence (15), we have:

$$\mathbb{E}_{\xi_{(t)}}[f(\widetilde{\mathbf{x}}^{(t+1)})] = \mathbb{E}_{\xi_{(t)}}f\left(\widetilde{\mathbf{x}}^{(t)} - \frac{\eta}{n(1-\beta)}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)})\right)$$
$$\overset{(a)}{\leq} f(\widetilde{\mathbf{x}}^{(t)}) \underbrace{-\left\langle\nabla f(\widetilde{\mathbf{x}}^{(t)}), \frac{\eta}{(1-\beta)}\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})\right\rangle}_{=:\ P_1} + \frac{L}{2}\frac{\eta^2}{(1-\beta)^2}\underbrace{\mathbb{E}_{\xi_{(t)}}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)})\right\|^2}_{=:\ P_2}, \tag{16}$$

where (a) follows from the L-smoothness of f. We show the following bounds on P_1 and P_2 in Appendix C-A.

$$P_1 \leq -\frac{\eta\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\|^2}{2(1-\beta)} + \frac{\eta L^2}{2n(1-\beta)}\sum_{i=1}^{n}\|\widetilde{\mathbf{x}}^{(t)}-\mathbf{x}_i^{(t)}\|^2 \tag{17}$$

$$P_2 \leq \frac{\sigma^2}{n} + \frac{2(M^2+n)L^2}{n^2}\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)}-\widetilde{\mathbf{x}}^{(t)}\right\|_2^2 + \frac{2(M^2+n)}{n}\left(G^2+B^2\left\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\right\|_2^2\right) \tag{18}$$

Substituting the bounds (17) and (18) in (16), we get:

$$\mathbb{E}_{\xi_{(t)}}[f(\widetilde{\mathbf{x}}^{(t+1)})] \leq f(\widetilde{\mathbf{x}}^{(t)}) + \frac{\eta^2L}{2(1-\beta)^2}\Big(\frac{\sigma^2+2(M^2+n)G^2}{n}\Big) + \left(\frac{\eta L^2}{2n(1-\beta)}+\frac{\eta^2L^3(M^2+n)}{n^2(1-\beta)^2}\right)\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)}-\widetilde{\mathbf{x}}^{(t)}\right\|_2^2 - \left(\frac{\eta}{2(1-\beta)}-\frac{\eta^2L(M^2+n)B^2}{n(1-\beta)^2}\right)\left\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\right\|_2^2. \tag{19}$$

When $\eta \leq \frac{n(1-\beta)}{2L(M^2+n)}$, we get $\left(\frac{\eta L^2}{2n(1-\beta)}+\frac{\eta^2L^3(M^2+n)}{n^2(1-\beta)^2}\right) \leq \frac{\eta L^2}{n(1-\beta)}$; and when $\eta \leq \frac{n(1-\beta)}{4LB^2(M^2+n)}$, we get $\left(\frac{\eta}{2(1-\beta)}-\frac{\eta^2L(M^2+n)B^2}{n(1-\beta)^2}\right) \geq \frac{\eta}{4(1-\beta)}$. Therefore, when $\eta \leq \min\{\frac{n(1-\beta)}{2L(M^2+n)}, \frac{n(1-\beta)}{4LB^2(M^2+n)}\}$, we get

$$\mathbb{E}_{\xi_{(t)}}[f(\widetilde{\mathbf{x}}^{(t+1)})] \leq f(\widetilde{\mathbf{x}}^{(t)}) + \frac{\eta^2L}{2(1-\beta)^2}\Big(\frac{\sigma^2+2(M^2+n)G^2}{n}\Big)$$

$$+\frac{\eta L^2}{n(1-\beta)}\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)}-\widetilde{\mathbf{x}}^{(t)}\right\|_2^2-\frac{\eta}{4(1-\beta)}\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\|_2^2 \quad (20)$$

By Jensen's inequality and L-smoothness of f, we have $\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|_2^2 \leq 2\left\|\nabla f(\overline{\mathbf{x}}^{(t)})-\nabla f(\widetilde{\mathbf{x}}^{(t)})\right\|_2^2 + 2\left\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\right\|_2^2 \leq 2L^2\left\|\overline{\mathbf{x}}^{(t)}-\widetilde{\mathbf{x}}^{(t)}\right\|_2^2 + 2\left\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\right\|_2^2$. Rearranging this gives $\left\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\right\|_2^2 \geq \frac{1}{2}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|_2^2 - L^2\left\|\overline{\mathbf{x}}^{(t)}-\widetilde{\mathbf{x}}^{(t)}\right\|_2^2$. Substituting this in (20) and rearranging:

$$\begin{aligned}
&\frac{\eta}{8(1-\beta)}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|_2^2 \leq f(\widetilde{\mathbf{x}}^{(t)})-\mathbb{E}_{\xi_{(t)}}[f(\widetilde{\mathbf{x}}^{(t+1)})]\\
&+\frac{\eta^2 L}{2(1-\beta)^2}\Big(\frac{\sigma^2+2(M^2+n)G^2}{n}\Big)+\frac{\eta L^2}{4(1-\beta)}\|\overline{\mathbf{x}}^{(t)}-\widetilde{\mathbf{x}}^{(t)}\|_2^2\\
&\quad+\frac{\eta L^2}{n(1-\beta)}\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)}-\widetilde{\mathbf{x}}^{(t)}\right\|_2^2\\
&\leq f(\widetilde{\mathbf{x}}^{(t)})-\mathbb{E}_{\xi_{(t)}}[f(\widetilde{\mathbf{x}}^{(t+1)})]+\frac{\eta^2 L}{2(1-\beta)^2}\frac{\sigma^2+2(M^2+n)G^2}{n}\\
&+\frac{2\eta L^2}{n(1-\beta)}\sum_{i=1}^{n}\|\mathbf{x}_i^{(t)}-\overline{\mathbf{x}}^{(t)}\|_2^2+\frac{9\eta L^2}{4(1-\beta)}\|\overline{\mathbf{x}}^{(t)}-\widetilde{\mathbf{x}}^{(t)}\|_2^2 \quad (21)
\end{aligned}$$

Now we bound $\left\|\overline{\mathbf{x}}^{(t)}-\widetilde{\mathbf{x}}^{(t)}\right\|_2^2$ in the following lemma, which we prove in Appendix C-A in supplementary material:

Lemma 5. *Consider the deviation of the global average parameter $\overline{\mathbf{x}}^{(t)}$ and the virtual sequence $\widetilde{\mathbf{x}}^{(t)}$ defined in* (13) *for constant stepsize η. Then at any time step t, we have:*

$$\|\overline{\mathbf{x}}^{(t)}-\widetilde{\mathbf{x}}^{(t)}\|^2 \leq \frac{\beta^4\eta^2}{(1-\beta)^3}\sum_{\tau=0}^{t-1}\beta^{t-\tau-1}\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(\tau)},\xi_i^{(\tau)})\|^2$$

Substituting the bound from Lemma 5 into (21) and then taking the expectation w.r.t. the entire past and average over $t=0$ to $t=T-1$ gives

$$\begin{aligned}
&\frac{\eta}{8T(1-\beta)}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|_2^2 \leq \frac{\eta^2 L}{2(1-\beta)^2}\frac{\sigma^2+2(M^2+n)G^2}{n}\\
&+\frac{1}{T}\mathbb{E}[f(\widetilde{\mathbf{x}}^{(0)})-f(\widetilde{\mathbf{x}}^{(T)})]+\sum_{t=0}^{T-1}\frac{2\eta L^2}{Tn(1-\beta)}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_i^{(t)}-\overline{\mathbf{x}}^{(t)}\|_2^2\\
&+\frac{9\eta^3\beta^4L^2}{4T(1-\beta)^4}\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\beta^{t-\tau-1}\mathbb{E}\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(\tau)},\xi_i^{(\tau)})\|^2
\end{aligned} \quad (22)$$

In the following lemma (which we prove in Appendix C-A) we bound the last term of (22).

Lemma 6. *Under setting of Theorem 1, it follows that:*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\left[\beta^{t-\tau-1}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(\tau)},\xi_i^{(\tau)})\right\|^2\right] \leq \frac{\sigma^2}{n(1-\beta)}\\
&+\frac{2(M^2+n)}{n(1-\beta)}\left(G^2+\frac{L^2}{T}\sum_{\tau=0}^{T-2}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}_i^{(\tau)}-\overline{\mathbf{x}}^{(\tau)}\right\|_2^2\right)\\
&+\frac{2(M^2+n)B^2}{n(1-\beta)}\frac{1}{T}\sum_{\tau=0}^{T-2}\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(\tau)})\|_2^2.
\end{aligned} \quad (23)$$

Substituting the bound from (23) into (120) and noting that $\widetilde{\mathbf{x}}^{(0)}=\overline{\mathbf{x}}^{(0)}$ and $f(\widetilde{\mathbf{x}}^{(T)}) \geq f^*$, where $f^*=f(\mathbf{x}^*)$, we get

$$\begin{aligned}
&\frac{\eta}{8(1-\beta)}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|_2^2 \leq \frac{f(\overline{\mathbf{x}}^{(0)})-f^*}{T}+\frac{\eta^2\sigma^2 L}{2n(1-\beta)^2}\\
&+\frac{\eta^2 L(M^2+n)G^2}{n(1-\beta)^2}+\frac{2\eta L^2}{n(1-\beta)}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}_i^{(t)}-\overline{\mathbf{x}}^{(t)}\right\|_2^2\\
&+\frac{9\eta^3\beta^4L^4(M^2+n)}{2(1-\beta)^5n^2}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}_i^{(t)}-\overline{\mathbf{x}}^{(t)}\right\|_2^2+\frac{9\eta^3\beta^4L^2\sigma^2}{4n(1-\beta)^5}\\
&+\frac{9\eta^3\beta^4L^2(M^2+n)}{2n(1-\beta)^5}\left(G^2+\frac{B^2}{T}\sum_{\tau=0}^{T-1}\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(\tau)})\|_2^2\right)\\
&=\frac{f(\overline{\mathbf{x}}^{(0)})-f^*}{T}+\frac{\eta^2 L}{2(1-\beta)^2}\Big(\frac{\sigma^2+2(M^2+n)G^2}{n}\Big)\Big(1+\frac{9\eta\beta^4}{2(1-\beta)^3}\Big)\\
&+\left(\frac{2\eta L^2}{n(1-\beta)}+\frac{9\eta^3\beta^4L^4(M^2+n)}{2n^2(1-\beta)^5}\right)\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_i^{(t)}-\overline{\mathbf{x}}^{(t)}\|_2^2\\
&\quad+\frac{9\eta^3\beta^4L^2(M^2+n)B^2}{2n(1-\beta)^5}\frac{1}{T}\sum_{\tau=0}^{T-1}\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(\tau)})\|_2^2
\end{aligned} \quad (24)$$

Note that (i) when $\eta \leq \frac{2(1-\beta)^3}{9\beta^4}$, we have $\left(1+\frac{9\eta\beta^4}{2(1-\beta)^3}\right) \leq 2$; (ii) when $\eta \leq \frac{2(1-\beta)^2}{3\beta^2 L}\sqrt{\frac{n}{M^2+n}}$, we have $\left(\frac{2\eta L^2}{n(1-\beta)}+\frac{9\eta^3\beta^4L^2}{4(1-\beta)^4}\frac{2(M^2+n)L^2}{n^2(1-\beta)}\right) \leq \frac{4\eta L^2}{n(1-\beta)}$; and (iii) when $\eta \leq \frac{(1-\beta)^2}{6\beta^2 LB}\sqrt{\frac{n}{2(M^2+n)}}$, we have $\frac{9\eta^3\beta^4L^2}{4(1-\beta)^4}\frac{2(M^2+n)B^2}{n(1-\beta)} \leq \frac{\eta}{16(1-\beta)}$. So, when $\eta \leq \min\{\frac{2(1-\beta)^3}{9\beta^4}, \frac{2(1-\beta)^2}{3\beta^2 L}\sqrt{\frac{n}{M^2+n}}, \frac{(1-\beta)^2}{6\beta^2 LB}\sqrt{\frac{n}{2(M^2+n)}}\}$, we get

$$\begin{aligned}
&\frac{\eta}{8(1-\beta)}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|_2^2 \leq \frac{f(\overline{\mathbf{x}}^{(0)})-f^*}{T}+\frac{\eta^2\sigma^2 L}{n(1-\beta)^2}\\
&+\frac{2(M^2+n)G^2\eta^2 L}{n(1-\beta)^2}+\frac{\eta}{16(1-\beta)}\frac{1}{T}\sum_{\tau=0}^{T-1}\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(\tau)})\|_2^2\\
&\quad+\frac{4\eta L^2}{n(1-\beta)}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}_i^{(t)}-\overline{\mathbf{x}}^{(t)}\right\|_2^2
\end{aligned} \quad (25)$$

Taking $\frac{\eta}{16(1-\beta)}\frac{1}{T}\sum_{\tau=0}^{T-1}\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(\tau)})\|_2^2$ to the LHS and multiplying both sides by $\frac{16(1-\beta)}{\eta}$ gives

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|_2^2 \leq \frac{16(1-\beta)(f(\overline{\mathbf{x}}^{(0)})-f^*)}{\eta T}\\
&+\frac{16\eta L}{(1-\beta)}\Big(\frac{\sigma^2+2(M^2+n)G^2}{n}\Big)+\frac{64L^2}{nT}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\mathbb{E}\|\mathbf{x}_i^{(t)}-\overline{\mathbf{x}}^{(t)}\|_2^2
\end{aligned} \quad (26)$$

D. Useful Lemmas

The following two lemmas (which we prove in Appendix C-B) will be useful for proving Lemma 2 and Lemma 3.

Lemma 7. *Under the setting of Theorem 1, for any $m \in \mathbb{N}$:*

$$\mathbb{E}\left\|\mathbf{X}^{((m+1)H)}-\overline{\mathbf{X}}^{((m+1)H)}\right\|_F^2 \leq a_1\mathbb{E}\left\|\mathbf{X}^{(mH)}-\overline{\mathbf{X}}^{(mH)}\right\|_F^2$$

$$+ a_2\mathbb{E}\left\|\mathbf{X}^{(mH)} - \widehat{\mathbf{X}}^{(mH)}\right\|_F^2 + a_3\eta^2\mathbb{E}\left\|\sum_{t'=mH}^{(m+1)H-1}\beta\mathbf{V}^{(t')} + \nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right\|_F^2, \quad (27)$$

where $a_1 = (1+\alpha_5^{-1})R_1$, $a_2 = (1+\alpha_5^{-1})R_2(1+\tau_1)(1-\omega)(1+\tau_2)$, *and* $a_3 = (R_1+R_2)(1+\alpha_5) + (1+\alpha_5^{-1})R_2\left((1+\tau_1^{-1}) + (1+\tau_1)(1-\omega)(1+\tau_2^{-1})\right)$. *Here,* $\tau_1,\tau_2,\alpha_5 > 0$ *are arbitrary numbers,* $R_1 = (1+\alpha_1)(1-\gamma\delta)^2$, $R_2 = (1+\alpha_1^{-1})\gamma^2\lambda^2$, $\alpha_1 > 0$, δ *is the spectral gap,* H *is synchronization gap,* γ *is consensus step-size,* $\lambda := \|\mathbf{W}-\mathbf{I}\|_2$ *where* $\mathbf{W}$ *is a doubly stochastic mixing matrix.*

Lemma 8. *Under the setting of Theorem 1, for any* $m \in \mathbb{N}$*:*

$$\mathbb{E}\|\mathbf{X}^{((m+1)H)} - \widehat{\mathbf{X}}^{((m+1)H)}\|_F^2 \le b_1\mathbb{E}\|\mathbf{X}^{(mH)} - \overline{\mathbf{X}}^{(mH)}\|_F^2 + b_2\mathbb{E}\|\mathbf{X}^{(mH)} - \widehat{\mathbf{X}}^{(mH)}\|_F^2 + b_3\eta^2\mathbb{E}\left\|\sum_{t'=mH}^{(m+1)H-1}\beta\mathbf{V}^{(t')} + \nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right\|_F^2, \quad (28)$$

where $b_1 = (1+\tau_3^{-1})\gamma^2\lambda^2(1+\tau_5)(1+\tau_6)$, $b_2 = (1+\tau_3)(1-\omega)(1+\tau_4) + (1+\tau_3^{-1})\gamma^2\lambda^2(1+\tau_5)(1+\tau_6^{-1})(1+\tau_7)(1-\omega)(1+\tau_8)$, $b_3 = (1+\tau_3)(1-\omega)(1+\tau_4^{-1}) + (1+\tau_3^{-1})\gamma^2\lambda^2(1+\tau_5)(1+\tau_6^{-1})\left((1+\tau_7^{-1}) + (1+\tau_7)(1-\omega)(1+\tau_8^{-1})\right) + (1+\tau_3^{-1})\gamma^2\lambda^2(1+\tau_5^{-1})$. *Here,* $\tau_3,\tau_4,\tau_5,\tau_6,\tau_7,\tau_8 > 0$ *are free parameters.*

E. Proof of Lemma 2

For any $t \in [T]$, define $m \in \lfloor \frac{t}{H} \rfloor - 1$. This implies that $(m+1)H \le t < (m+2)H$. Now we note that:

$$\begin{aligned}
\Xi^{(t)} &:= \mathbb{E}\left\|\mathbf{X}^{(t)} - \overline{\mathbf{X}}^{(t)}\right\|_F^2 \\
&= \mathbb{E}\left\|\mathbf{X}^{(t)} - \overline{\mathbf{X}}^{((m+1)H)} - \left(\overline{\mathbf{X}}^{(t)} - \overline{\mathbf{X}}^{((m+1)H)}\right)\right\|_F^2 \\
&\overset{(a)}{\le} \mathbb{E}\left\|\mathbf{X}^{(t)} - \overline{\mathbf{X}}^{((m+1)H)}\right\|_F^2 \quad (29) \\
&\le (1+\nu_1)\mathbb{E}\left\|\mathbf{X}^{((m+1)H)} - \overline{\mathbf{X}}^{((m+1)H)}\right\|_F^2 \\
&\quad + (1+\nu_1^{-1})\eta^2\mathbb{E}\left\|\sum_{t'=(m+1)H}^{t-1}\left(\beta\mathbf{V}^{(t')} + \nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right)\right\|_F^2 \\
&\overset{(b)}{\le} (1+\nu_1)(a_1\Xi^{(mH)} + a_2\mathbb{E}\|\mathbf{X}^{(mH)} - \widehat{\mathbf{X}}^{(mH)}\|_F^2) \\
&\quad + (1+\nu_1)a_3\eta^2\mathbb{E}\left\|\sum_{t'=mH}^{(m+1)H-1}\beta\mathbf{V}^{(t')} + \nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right\|_F^2 \\
&\quad + (1+\nu_1^{-1})\eta^2\mathbb{E}\left\|\sum_{t'=(m+1)H}^{t-1}\beta\mathbf{V}^{(t')} + \nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right\|_F^2 \\
&\le (1+\nu_1)(a_1\Xi^{(mH)} + a_2\mathbb{E}\|\mathbf{X}^{(mH)} - \widehat{\mathbf{X}}^{(mH)}\|_F^2) \\
&\quad + (1+\nu_1)a_3\eta^2 H\sum_{t'=mH}^{t-1}\mathbb{E}\|\beta\mathbf{V}^{(t')} + \nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\|_F^2 \\
&\quad + (1+\nu_1^{-1})\eta^2 H\sum_{t'=mH}^{t-1}\mathbb{E}\|\beta\mathbf{V}^{(t')} + \nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\|_F^2 \\
&\le (1+\nu_1)a_1\Xi^{(mH)} + (1+\nu_1)a_2\mathbb{E}\|\mathbf{X}^{(mH)} - \widehat{\mathbf{X}}^{(mH)}\|_F^2 \\
&\quad + 2\left((1+\nu_1)a_3 + (1+\nu_1^{-1})\right)\eta^2 H\sum_{t'=mH}^{t-1}\mathbb{E}\|\nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\|_F^2 \\
&\quad + 2\left((1+\nu_1)a_3 + (1+\nu_1^{-1})\right)\eta^2 H\sum_{t'=mH}^{t-1}\beta^2\mathbb{E}\left\|\mathbf{V}^{(t')}\right\|_F^2 \quad (30)
\end{aligned}$$

Here, (a) follows from the inequality: $\frac{1}{n}\sum_{i=1}^n\left\|\mathbf{a}_i - \frac{1}{n}\sum_{i=1}^n\mathbf{a}_i\right\|_2^2 \le \frac{1}{n}\sum_{i=1}^n\|\mathbf{a}_i\|_2^2$ and (b) follows from (27) (in Lemma 7). The coefficients a_1, a_2, a_3 in the RHS of (b) are defined in Lemma 7.

Proposition 3. *For any* t'*, we have:*

$$\mathbb{E}\left\|\nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right\|_F^2 \le 2(M^2+1)(L^2\Xi^{(t')} + nG^2) + 2(M^2+1)nB^2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|_2^2 + n\sigma^2 \quad (31)$$

Substituting (31) into (30), for $(m+1)H \le t < (m+2)H$:

$$\begin{aligned}
\Xi^{(t)} &\le (1+\nu_1)\left(a_1\Xi^{(mH)} + a_2\mathbb{E}\left\|\mathbf{X}^{(mH)} - \widehat{\mathbf{X}}^{(mH)}\right\|_F^2\right) \\
&\quad + 2c_2\eta^2H^2n\left(2(M^2+1)G^2 + \sigma^2\right) + c_2\eta^2H\beta^2\sum_{t'=mH}^{t-1}\mathbb{E}\|\mathbf{V}^{(t')}\|_F^2 \\
&\quad + 2c_2\eta^2H(M^2+1)\sum_{t'=mH}^{t-1}L^2\Xi^{(t')} + nB^2\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(t')})\|^2 \quad (32)
\end{aligned}$$

where $c_2 = 2\left((1+\nu_1)a_3 + (1+\nu_1^{-1})\right)$. For any $j \in [T]$ and $m' = \lfloor \frac{j}{H} \rfloor - 1$, define

$$S^{(j)} := \Xi^{(j)} + \mathbb{E}\left\|\mathbf{X}^{(j)} - \widehat{\mathbf{X}}^{((m'+1)H)}\right\|_F^2. \quad (33)$$

By definition, we have $S^{(mH)} = \Xi^{(mH)} + \mathbb{E}\left\|\mathbf{X}^{(mH)} - \widehat{\mathbf{X}}^{(mH)}\right\|_F^2$ and also that $\Xi^{(t')} \le S^{(t')}$ for any t'. Using these in (32), we get

$$\begin{aligned}
\Xi^{(t)} &\le (1+\nu_1)\left(a_1\Xi^{(mH)} + a_2\mathbb{E}\left\|\mathbf{X}^{(mH)} - \widehat{\mathbf{X}}^{(mH)}\right\|_F^2\right) \\
&\quad + 2c_2\eta^2H^2n\left(2(M^2+1)G^2 + \sigma^2\right) + c_2\eta^2H\beta^2\sum_{t'=mH}^{t-1}\mathbb{E}\|\mathbf{V}^{(t')}\|_F^2 \\
&\quad + 2c_2\eta^2H(M^2+1)\sum_{t'=mH}^{t-1}L^2S^{(t')} + nB^2\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(t')})\|_2^2 \quad (34)
\end{aligned}$$

Our aim is to get an upper-bound on $S^{(t)}$, which is defined in (33) as $S^{(t)} = \Xi^{(t)} + \mathbb{E}\left\|\mathbf{X}^{(t)} - \widehat{\mathbf{X}}^{(\lfloor t/H \rfloor H)}\right\|_F^2$. However, in (34), we have only derived an upper-bound on $\Xi^{(t)}$ in terms of $S^{(t')}$ for $t' < t$. So,, we need to derive a similar upper-bound on the other term $\mathbb{E}\left\|\mathbf{X}^{(t)} - \widehat{\mathbf{X}}^{(\lfloor t/H \rfloor H)}\right\|_F^2$, and then we will add both the upper-bounds to get an upper-bound on $S^{(t)}$. In the following, we derive an upper bound on $\mathbb{E}\left\|\mathbf{X}^{(t)} - \widehat{\mathbf{X}}^{(\lfloor t/H \rfloor H)}\right\|_F^2$. Let $m = \lfloor \frac{t}{H} \rfloor - 1$, we have:

$$\mathbb{E}\left\|\mathbf{X}^{(t)} - \widehat{\mathbf{X}}^{((m+1)H)}\right\|_F^2 = \mathbb{E}\left\|\mathbf{X}^{((m+1)H)} - \widehat{\mathbf{X}}^{((m+1)H)} - \eta\sum_{t'=(m+1)H}^{t-1}\left(\beta\mathbf{V}^{(t')} + \nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right)\right\|_F^2$$

$$\leq (1+\nu_1)\mathbb{E}\left\|\mathbf{X}^{((m+1)H)}-\widehat{\mathbf{X}}^{((m+1)H)}\right\|_F^2$$
$$+(1+\nu_1^{-1})\eta^2\mathbb{E}\left\|\sum_{t'=(m+1)H}^{t-1}\left(\beta\mathbf{V}^{(t')}+\nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right)\right\|_F^2$$
$$\overset{(a)}{\leq}(1+\nu_1)(b_1\Xi^{(mH)}+b_2\mathbb{E}\|\mathbf{X}^{(mH)}-\widehat{\mathbf{X}}^{(mH)}\|_F^2)$$
$$+(1+\nu_1)b_3\eta^2\mathbb{E}\left\|\sum_{t'=mH}^{(m+1)H-1}\left(\beta\mathbf{V}^{(t')}+\nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right)\right\|_F^2$$
$$+(1+\nu_1^{-1})\eta^2\mathbb{E}\left\|\sum_{t'=(m+1)H}^{t-1}\left(\beta\mathbf{V}^{(t')}+\nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right)\right\|_F^2$$
$$\leq(1+\nu_1)\left(b_1\Xi^{(mH)}+b_2\mathbb{E}\left\|\mathbf{X}^{(mH)}-\widehat{\mathbf{X}}^{(mH)}\right\|_F^2\right)$$
$$+2\left((1+\nu_1)b_3+(1+\nu_1^{-1})\right)\eta^2H\sum_{t'=mH}^{t-1}\beta^2\mathbb{E}\|\mathbf{V}^{(t')}\|_F^2$$
$$+2\left((1+\nu_1)b_3+(1+\nu_1^{-1})\right)\eta^2H\sum_{t'=mH}^{t-1}\mathbb{E}\|\nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\|_F^2$$
$$\overset{(b)}{\leq}(1+\nu_1)\left(b_1\Xi^{(mH)}+b_2\mathbb{E}\left\|\mathbf{X}^{(mH)}-\widehat{\mathbf{X}}^{(mH)}\right\|_F^2\right)$$
$$+2c_4\eta^2H^2n\left(2(M^2+1)G^2+\sigma^2\right)+c_4\eta^2H\beta^2\sum_{t'=mH}^{t-1}\mathbb{E}\|\mathbf{V}^{(t')}\|_F^2$$
$$+2c_4\eta^2H(M^2+1)\sum_{t'=mH}^{t-1}L^2\Xi^{(t')}+nB^2\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(t')})\|_2^2 \quad (35)$$

where (a) follows from (28) in Lemma 8 and the coefficients b_1, b_2, b_3 in the RHS of (a) are defined in Lemma 8, and (b) follows from substituting the bound from (31) (in Proposition 3). In the RHS of (b), $c_4 = 2\left((1+\nu_1)b_3+(1+\nu_1^{-1})\right)$. Adding (34), (35) for $S^{(t)} = \Xi^{(t)} + \mathbb{E}\|\mathbf{X}^{(t)} - \widehat{\mathbf{X}}^{((m+1)H)}\|_F^2$:

$$S^{(t)} \leq (1+\nu_1)\max\{a_1+b_1, a_2+b_2\}S^{(mH)} + 2c_1\eta^2H^2\Gamma$$
$$+c_1\eta^2H\beta^2\sum_{t'=mH}^{t-1}\mathbb{E}\|\mathbf{V}^{(t')}\|_F^2+2c_1\eta^2H(M^2+1)L^2\sum_{t'=mH}^{t-1}S^{(t')}$$
$$+2c_1\eta^2H(M^2+1)nB^2\textstyle\sum_{t'=mH}^{t-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|_2^2 \quad (36)$$

where $\Gamma = n\left(2(M^2+1)G^2+\sigma^2\right)$ and $c_1 = c_2 + c_4$ with $c_2 = 2\left((1+\nu_1)a_3+(1+\nu_1^{-1})\right)$ and $c_4 = 2\left((1+\nu_1)b_3+(1+\nu_1^{-1})\right)$. Here, $\nu_1 > 0$ is a free coefficient, and a_1, a_2, a_3 and b_1, b_2, b_3 are defined in Lemma 7 and Lemma 8, respectively. We will set the free variables such that the coefficients of $S^{(t')}$ for any $t' = mH, ..., t-1$ on the RHS become strictly less than one.

In Appendix C-C, we show that if we set the free parameters to be the following:

$$\tau_i = \frac{\omega}{4}, \text{ for } i = 1,2,3,4,5,7,8; \quad \tau_6 = \frac{4}{\omega}; \quad \nu_1 = \frac{\gamma^*\delta}{4};$$
$$\alpha_1 = \frac{\gamma\delta}{2}; \quad \alpha_5^{-1} = \frac{\gamma\delta}{2}; \quad \gamma = \frac{2\delta\omega^3}{(128\lambda^2+24\lambda^2\omega^2+4\delta^2\omega^2)};$$

Then we get

$$(1+\nu_1)\max\{a_1+b_1, a_2+b_2\} \leq 1-\frac{\gamma^*\delta}{4} \leq 1-\frac{\delta^2\omega^3}{1224}, \quad (37)$$

$$c_1 \leq 2(1+\frac{\gamma\delta}{4})\left(\frac{3}{\gamma\delta}+\frac{9\lambda^2}{\delta^2}+\frac{45\gamma\lambda^2}{\delta\omega}+\frac{104\gamma^2\lambda^2}{\omega^2}+\frac{4}{\omega}-2\right)$$
$$+4(1+\frac{4}{\gamma\delta}). \quad (38)$$

Putting these bounds back into (36), we get the following upper bound for $(m+1)H \leq t \leq (m+2)H-1$:

$$S^{(t)} \leq \left(1-\frac{\gamma\delta}{4}\right)S^{(mH)} + 2c_1\eta^2H^2n\left(2(M^2+1)G^2+\sigma^2\right)$$
$$+c_1\eta^2H\beta^2\sum_{t'=mH}^{t-1}\mathbb{E}\|\mathbf{V}^{(t')}\|_F^2+2c_1\eta^2H(M^2+1)L^2\sum_{t'=mH}^{t-1}S^{(t')}$$
$$+2c_1\eta^2H(M^2+1)nB^2\sum_{t'=mH}^{t-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|_2^2. \quad (39)$$

F. Proof of Lemma 3

For any fixed $t \in [T]$ and the corresponding $m \in \lfloor \frac{t}{H} \rfloor - 1$, in Section V-E, we derived an upper-bound on $S^{(\hat{t})}$ all $\hat{t} \in [T]$ such that $(m+1)H \leq \hat{t} < (m+2)H$ (note that t and $\hat{t}$ will give exactly the same terms in Section V-E, so we just kept t everywhere). In this section, we consider the case when $mH \leq \hat{t} < (m+1)H$.

$$\Xi^{(\hat{t})} \overset{(a)}{\leq} \mathbb{E}\left\|\mathbf{X}^{(\hat{t})}-\overline{\mathbf{X}}^{(mH)}\right\|_F^2 \quad (40)$$
$$\leq (1+\nu_3)\mathbb{E}\left\|\mathbf{X}^{(mH)}-\overline{\mathbf{X}}^{(mH)}\right\|_F^2$$
$$+(1+\nu_3^{-1})\eta^2\mathbb{E}\left\|\sum_{t'=mH}^{\hat{t}-1}\left(\beta\mathbf{V}^{(t')}+\nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right)\right\|_F^2$$
$$\overset{(b)}{\leq}(1+\nu_3)\Xi^{(mH)}+2(1+\nu_3^{-1})\eta^2H\beta^2\sum_{t'=mH}^{\hat{t}-1}\mathbb{E}\left\|\mathbf{V}^{(t')}\right\|_F^2$$
$$+2(1+\nu_3^{-1})\eta^2H\sum_{t'=mH}^{\hat{t}-1}\mathbb{E}\left\|\nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right\|_F^2$$
$$\overset{(c)}{\leq}(1+\nu_3)\Xi^{(mH)}+2(1+\nu_3^{-1})\eta^2H\beta^2\sum_{t'=mH}^{\hat{t}-1}\mathbb{E}\left\|\mathbf{V}^{(t')}\right\|_F^2$$
$$+4(M^2+1)(1+\nu_3^{-1})\eta^2H\textstyle\sum_{t'=mH}^{\hat{t}-1}\left(L^2\Xi^{(t')}+nG^2\right)$$
$$+2(1+\nu_3^{-1})\eta^2H\sum_{t'=mH}^{\hat{t}-1}\left(2(M^2+1)nB^2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|_2^2+n\sigma^2\right)$$
$$\leq(1+\nu_3)\Xi^{(mH)}+2(1+\nu_3^{-1})\eta^2H^2n\left(2(M^2+1)G^2+\sigma^2\right)$$
$$+4(1+\nu_3^{-1})\eta^2H(M^2+1)\sum_{t'=mH}^{\hat{t}-1}L^2\Xi^{(t')}+nB^2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|_2^2$$
$$+2(1+\nu_3^{-1})\eta^2H\beta^2\textstyle\sum_{t'=mH}^{\hat{t}-1}\mathbb{E}\left\|\mathbf{V}^{(t')}\right\|_F^2 \quad (41)$$

where (a) follows from the same reasoning using which we obtained (29), (b) uses $\Xi^{(mH)} = \mathbb{E}\left\|\mathbf{X}^{(mH)}-\overline{\mathbf{X}}^{(mH)}\right\|_F^2$, and (c) follows from (31) (in Proposition 3).

As mentioned in Section V-E, our aim is to get an upper-bound on $S^{(\hat{t})}$, which is defined in (33) as $S^{(\hat{t})} = \Xi^{(\hat{t})} +$

$\mathbb{E}\left\|\mathbf{X}^{(\hat{t})}-\widehat{\mathbf{X}}^{(\lfloor \hat{t}/H\rfloor H)}\right\|_F^2$. However, in (41), we have only derived an upper-bound on $\Xi^{(\hat{t})}$. So, we need to derive a similar upper-bound on the other term $\mathbb{E}\left\|\mathbf{X}^{(\hat{t})}-\widehat{\mathbf{X}}^{(\lfloor \hat{t}/H\rfloor H)}\right\|_F^2$, and then adding both the upper-bounds gives a bound on $S^{(\hat{t})}$. Note that since $mH \le \hat{t} < (m+1)H$, we have $\lfloor \frac{\hat{t}}{H} \rfloor = m$. In order to upper-bound $\mathbb{E}\left\|\mathbf{X}^{(\hat{t})}-\widehat{\mathbf{X}}^{(mH)}\right\|_F^2$, we can follow the same steps that we used from (40) to (41) (just replace $\overline{\mathbf{X}}^{(mH)}$ with $\widehat{\mathbf{X}}^{(mH)}$). This would give the following bound:

$$\begin{aligned}
&\mathbb{E}\left\|\mathbf{X}^{(\hat{t})}-\widehat{\mathbf{X}}^{(mH)}\right\|_F^2 \le (1+\nu_3)\mathbb{E}\left\|\mathbf{X}^{(mH)}-\widehat{\mathbf{X}}^{(mH)}\right\|_F^2 \\
&+2(1+\nu_3^{-1})\eta^2H^2[n\left(2(M^2+1)G^2+\sigma^2\right)+\beta^2\sum_{t'=mH}^{\hat{t}-1}\mathbb{E}\|\mathbf{V}^{(t')}\|_F^2] \\
&+4(1+\nu_3^{-1})\eta^2H(M^2+1)nB^2\sum_{t'=mH}^{\hat{t}-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|_2^2 \\
&+4(1+\nu_3^{-1})\eta^2H(M^2+1)L^2\sum_{t'=mH}^{\hat{t}-1}\Xi^{(t')} \qquad (42)
\end{aligned}$$

Adding (41) and (42), and using the definition that $S^{(\hat{t})} = \Xi^{(\hat{t})} + \mathbb{E}\left\|\mathbf{X}^{(\hat{t})}-\widehat{\mathbf{X}}^{(\lfloor \hat{t}/H\rfloor H)}\right\|_F^2$ together with that $\Xi^{(t')} \le S^{(t')}$, and taking $\nu_3 = \frac{\gamma\delta}{4}$, we get:

$$\begin{aligned}
&S^{(\hat{t})} \le (1+\frac{\gamma\delta}{4})S^{(mH)}+4(1+\frac{4}{\gamma\delta})\eta^2H^2n\left(2(M^2+1)G^2+\sigma^2\right) \\
&+8(1+\frac{4}{\gamma\delta})\eta^2H(M^2+1)\sum_{t'=mH}^{\hat{t}-1}(L^2S^{(t')}+nB^2\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(t')})\|_2^2) \\
&+4(1+\frac{4}{\gamma\delta})\eta^2H\beta^2\sum_{t'=mH}^{\hat{t}-1}\mathbb{E}\|\mathbf{V}^{(t')}\|_F^2 \qquad (43)
\end{aligned}$$

In order to make our calculations less cluttered later, we would like to write all terms (except the first one) in the RHS above in the same form as given in (39). Indeed, it can be verified easily that $4(1+\frac{4}{\gamma\delta}) \le c_1$, where c_1 is exactly the same as in (39). Substituting this in (43) above yields the bound below for $mH \le \hat{t} < (m+1)H$, where $m \in \lfloor \frac{t}{H} \rfloor - 1$:

$$\begin{aligned}
&S^{(\hat{t})} \le (1+\frac{\gamma\delta}{4})S^{(mH)}+2c_1\eta^2H^2n\left(2(M^2+1)G^2+\sigma^2\right) \\
&+2c_1\eta^2H(M^2+1)\sum_{t'=mH}^{\hat{t}-1}(L^2S^{(t')}+nB^2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|_2^2) \\
&+c_1\eta^2H\beta^2\sum_{t'=mH}^{\hat{t}-1}\mathbb{E}\left\|\mathbf{V}^{(t')}\right\|_F^2 \qquad (44)
\end{aligned}$$

where c_1 is exactly the same as in (39).

G. Proof of Lemma 4

Let $A = 2c_1H^2n\left(2(M^2+1)G^2+\sigma^2\right)$, $D = \frac{c_1H\beta^2}{(1-\beta)}$, $C = 2c_1H(M^2+1)nB^2$, and $\Lambda^{(t')} = (1-\beta)\mathbb{E}\left\|\mathbf{V}^{(t)}\right\|_F^2$, where c_1 is the same as in (104). Since $\eta \le \sqrt{\frac{\gamma\delta}{512c_1H^2(M^2+1)L^2}}$, we have $2c_1\eta^2H(M^2+1)L^2 \le \frac{\gamma\delta}{4}\frac{1}{64H}$.

Take any $t \in [T]$ and let $m = \lfloor \frac{t}{H} \rfloor - 1$. With these substitutions and letting $\alpha = \frac{\gamma\delta}{4}$, the bound from (39) for any t such that $(m+1)H \le t \le (m+2)H-1$ becomes:

$$\begin{aligned}
S^{(t)} &\le \left(1-\frac{\alpha}{2}\right)S^{(mH)}+A\eta^2+\frac{\alpha}{64H}\sum_{t'=mH}^{t-1}S^{(t')} \\
&+C\eta^2\sum_{t'=mH}^{t-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|^2+D\eta^2\sum_{t'=mH}^{t-1}\Lambda^{(t')}. \qquad (45)
\end{aligned}$$

And for any $\hat{t}$ such that $mH \le \hat{t} < (m+1)H$, the bound from (44) becomes:

$$\begin{aligned}
S^{(\hat{t})} &\le \left(1-\frac{\alpha}{2}\right)S^{(mH)}+A\eta^2+\frac{\alpha}{64H}\sum_{t'=mH}^{\hat{t}-1}S^{(t')} \\
&+C\eta^2\sum_{t'=mH}^{\hat{t}-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|^2+D\eta^2\sum_{t'=mH}^{\hat{t}-1}\Lambda^{(t')}. \qquad (46)
\end{aligned}$$

Consider (45). Substituting the value of $S^{(t-1)}$ recursively in the RHS of (45), we get:

$$\begin{aligned}
S^{(t)} &\le \left(1-\frac{\alpha}{2}\right)S^{(mH)}+A\eta^2+\frac{\alpha}{64H}\sum_{t'=mH}^{t-2}S^{(t')} \\
&+C\eta^2\sum_{t'=mH}^{t-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|^2+D\eta^2\sum_{t'=mH}^{t-1}\Lambda^{(t')} \\
&+\frac{\alpha}{64H}\Bigg(\left(1-\frac{\alpha}{2}\right)S^{(mH)}+A\eta^2+\frac{\alpha}{64H}\sum_{t'=mH}^{t-2}S^{(t')} \\
&\quad +C\eta^2\sum_{t'=mH}^{t-2}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|^2+D\eta^2\sum_{t'=mH}^{t-2}\Lambda^{(t')}\Bigg) \\
&=\left(1-\frac{\alpha}{2}\right)\left(1+\frac{\alpha}{64H}\right)S^{(mH)}+A\left(1+\frac{\alpha}{64H}\right)\eta^2+D\eta^2\Lambda_{t-1} \\
&+\frac{\alpha}{64H}\left(1+\frac{\alpha}{64H}\right)\sum_{t'=mH}^{t-2}S^{(t')}+\left(1+\frac{\alpha}{64H}\right)D\eta^2\sum_{t'=mH}^{t-2}\Lambda^{(t')} \\
&+\left(1+\frac{\alpha}{64H}\right)C\eta^2\sum_{t'=mH}^{t-2}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|^2+C\eta^2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t-1)})\right\|^2
\end{aligned}$$

Substituting the values in the RHS till $(m+1)H$, we get:

$$\begin{aligned}
S^{(t)} &\le \left(1-\frac{\alpha}{2}\right)\left(1+\frac{\alpha}{64H}\right)^H S^{(mH)}+A\left(1+\frac{\alpha}{64H}\right)^H\eta^2 \\
&+\frac{\alpha}{64H}\left(1+\frac{\alpha}{64H}\right)^H\sum_{t'=mH}^{(m+1)H-1}S^{(t')} \\
&+\left(1+\frac{\alpha}{64H}\right)^H\eta^2\sum_{t'=mH}^{(m+1)H-1}(C\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|^2+D\Lambda^{(t')}) \\
&+\eta^2\sum_{t'=(m+1)H}^{t-1}\left(1+\frac{\alpha}{64H}\right)^{t-1-t'}(C\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(t')})\|^2+D\Lambda^{(t')})
\end{aligned}$$

Now consider t' such that $mH \le t' < (m+1)H$. Substituting the value of $S^{((m+1)H-1)}$ from (46) int the R.H.S above gives:

$$S^{(t)} \le \left(1-\frac{\alpha}{2}\right)\left(1+\frac{\alpha}{64H}\right)^H S^{(mH)}+A\left(1+\frac{\alpha}{64H}\right)^H\eta^2$$

$$+\frac{\alpha}{64H}\left(1+\frac{\alpha}{64H}\right)^{H}\sum_{t'=mH}^{(m+1)H-2}S^{(t')}$$
$$+\left(1+\frac{\alpha}{64H}\right)^{H}\eta^2\sum_{t'=mH}^{(m+1)H-1}(C\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|^2+D\Lambda^{(t')})$$
$$+\frac{\alpha}{64H}\left(1+\frac{\alpha}{64H}\right)^{H}\Bigg[(1+\frac{\alpha}{2})S^{(mH)}+\frac{\alpha}{64H}\sum_{j=mH}^{(m+1)H-2}S^{(j)}$$
$$+C\eta^2\sum_{j=mH}^{(m+1)H-2}\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(j)})\|^2+D\eta^2\sum_{j=mH}^{(m+1)H-2}\Lambda^{(j)}+A\eta^2\Bigg]$$
$$+\eta^2\sum_{t'=(m+1)H}^{t-1}\left(1+\frac{\alpha}{64H}\right)^{t-1-t'}(C\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(t')})\|^2+D\Lambda^{(t')})$$
$$\leq\left(\left(1-\frac{\alpha}{2}\right)+\frac{\alpha}{64H}\left(1+\frac{\alpha}{2}\right)\right)\left(1+\frac{\alpha}{64H}\right)^{H}S^{(mH)}$$
$$+A\left(1+\frac{\alpha}{64H}\right)^{H+1}\eta^2+\frac{\alpha}{64H}\left(1+\frac{\alpha}{64H}\right)^{H+1}\sum_{t'=mH}^{(m+1)H-2}S^{(t')}$$
$$+\left(1+\frac{\alpha}{64H}\right)^{H+1}\eta^2\sum_{t'=mH}^{(m+1)H-2}(C\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(t')})\|^2+D\Lambda^{(t')})$$
$$+\eta^2\sum_{t'=(m+1)H}^{t-1}\left(1+\frac{\alpha}{64H}\right)^{t-1-t'}(C\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|^2+D\Lambda^{(t')})$$
$$+\eta^2\left(1+\frac{\alpha}{64H}\right)^{H}(C\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{((m+1)H-1)})\|^2+D\Lambda^{((m+1)H-1)})$$

Now we note that for $0<\alpha\leq 1$, $\frac{\alpha}{64H}\left(1+\frac{\alpha}{2}\right)\leq\left(1-\frac{\alpha}{2}\right)\frac{\alpha}{16H}$. Using this fact in the first term and $\left(1+\frac{\alpha}{64H}\right)\leq\left(1+\frac{\alpha}{16H}\right)$, and $\left(1+\frac{\alpha}{64H}\right)^{t-1-t'}\leq\left(1+\frac{\alpha}{16H}\right)^{H}$ for all $t'\in\{(m+1)H,\ldots,t-1\}$ in the RHS above gives:

$$S^{(t)}\leq\left(1-\frac{\alpha}{2}\right)\left(1+\frac{\alpha}{16H}\right)^{H+1}S^{(mH)}+A\left(1+\frac{\alpha}{16H}\right)^{H+1}\eta^2$$
$$+\frac{\alpha}{64H}\left(1+\frac{\alpha}{16H}\right)^{H+1}\sum_{t'=mH}^{(m+1)H-2}S^{(t')}$$
$$+\left(1+\frac{\alpha}{16H}\right)^{H+1}\eta^2\sum_{t'=mH}^{(m+1)H-2}(C\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|^2+D\Lambda^{(t')})$$
$$+\eta^2\left(1+\frac{\alpha}{16H}\right)^{H}\sum_{t'=(m+1)H}^{t-1}(C\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|^2+D\Lambda^{(t')})$$
$$+\eta^2\left(1+\frac{\alpha}{16H}\right)^{H}(C\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{((m+1)H-1)})\right\|^2+D\Lambda^{((m+1)H-1)})$$

Using $\left(1+\frac{\alpha}{16H}\right)^{H}\leq\left(1+\frac{\alpha}{16H}\right)^{H+1}$ in the last two terms and then clubbing together terms respectively with C and D:

$$S^{(t)}\leq\left(1-\frac{\alpha}{2}\right)\left(1+\frac{\alpha}{16H}\right)^{H+1}S^{(mH)}+A\left(1+\frac{\alpha}{16H}\right)^{H+1}\eta^2$$
$$+\left(1+\frac{\alpha}{16H}\right)^{H+1}\eta^2\sum_{t'=mH}^{t-1}(C\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|^2+D\Lambda^{(t')})$$
$$+\frac{\alpha}{64H}\left(1+\frac{\alpha}{16H}\right)^{H+1}\sum_{t'=mH}^{(m+1)H-2}S^{(t')}$$

Recursively substituting the values till mH gives us:

$$S^{(t)}\leq\left(1-\frac{\alpha}{2}\right)\left(1+\frac{\alpha}{16H}\right)^{2H}S^{(mH)}+A\left(1+\frac{\alpha}{16H}\right)^{2H}\eta^2$$
$$+\left(1+\frac{\alpha}{16H}\right)^{2H}\eta^2\sum_{t'=mH}^{t-1}(C\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|^2+D\Lambda^{(t')})$$

For $\alpha\leq 1$, we note that $\left(1+\frac{\alpha}{16H}\right)^{2H}\leq e^{\frac{\alpha}{8}}\leq 1+\frac{\alpha}{4}$. Plugging this in the first term on the RHS and using $\left(1-\frac{\alpha}{2}\right)\left(1+\frac{\alpha}{4}\right)\leq\left(1-\frac{\alpha}{4}\right)$ and $\left(1+\frac{\alpha}{16H}\right)^{2H}\leq 1+\frac{\alpha}{4}\leq 2$ gives us the following recursion equation for any $t\in[T]$:

$$S^{(t)}\leq\left(1-\frac{\alpha}{4}\right)S^{(mH)}+2A\eta^2$$
$$+2C\eta^2\sum_{t'=mH}^{t-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|^2+2D\eta^2\sum_{t'=mH}^{t-1}\Lambda^{(t')}\quad(47)$$

Unrolling recursion equation in (47) for $S^{(mH)}$ till 0, we get:

$$S^{(t)}\leq 2A\eta^2\sum_{j=0}^{m-1}\left(1-\frac{\alpha}{4}\right)^{j}+2D\eta^2\sum_{j=0}^{t-1}\left(1-\frac{\alpha}{4}\right)^{\lfloor\frac{t-j}{H}\rfloor}\Lambda^{(j)}$$
$$+2C\eta^2\sum_{j=0}^{t-1}\left(1-\frac{\alpha}{4}\right)^{\lfloor\frac{t-j}{H}\rfloor}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(j)})\right\|^2\quad(48)$$

Note that $\sum_{j=0}^{m-1}\left(1-\frac{\alpha}{4}\right)^{j}\leq\frac{4}{\alpha}$. Using this and the bound $\left(1-\frac{\alpha}{4}\right)^{\lfloor\frac{t-j}{H}\rfloor}\leq 2\left(1-\frac{\alpha}{8H}\right)^{t-j}$ (proved in Appendix C-E) into (48) gives us:

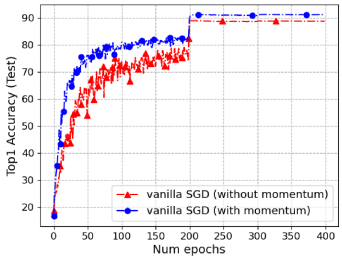
$$S^{(t)}\leq\frac{8A\eta^2}{\alpha}+4C\eta^2\sum_{j=0}^{t-1}\left(1-\frac{\alpha}{8H}\right)^{t-j}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(j)})\right\|^2$$
$$+4D\eta^2\sum_{j=0}^{t-1}\left(1-\frac{\alpha}{8H}\right)^{t-j}\Lambda^{(j)}$$

Taking summation from $t=0$ to $T-1$, we get:

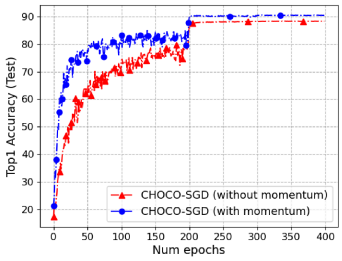
$$\sum_{t=0}^{T-1}S^{(t)}\leq 4C\eta^2\sum_{t=0}^{T-1}\sum_{j=0}^{t-1}\left(1-\frac{\alpha}{8H}\right)^{t-j}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(j)})\right\|^2$$
$$+4D\eta^2\sum_{t=0}^{T-1}\sum_{j=0}^{t-1}\left(1-\frac{\alpha}{8H}\right)^{t-j}\Lambda^{(j)}+\frac{8A\eta^2}{\alpha}T$$
$$\leq\frac{8A\eta^2}{\alpha}T+4C\eta^2\sum_{j=0}^{T-1}\sum_{t=j+1}^{T-1}\left(1-\frac{\alpha}{8H}\right)^{t-j}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(j)})\right\|^2$$
$$+4D\eta^2\sum_{j=0}^{T-1}\sum_{t=j+1}^{T-1}\left(1-\frac{\alpha}{8H}\right)^{t-j}\Lambda^{(j)}$$
$$\leq\frac{8A\eta^2T}{\alpha}+\frac{32C\eta^2H}{\alpha}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(t)})\|^2+\frac{32DH\eta^2}{\alpha}\sum_{t=0}^{T-1}\Lambda^{(t')}\quad(49)$$

To bound the last term in the RHS of (49), from the definition of $\Lambda^{(t')}$ in (12), note that:

$$\sum_{t=0}^{T-1}\Lambda^{(t')}=\sum_{t=0}^{T-1}\sum_{j=0}^{t}\beta^{t-j}\mathbb{E}\left\|\nabla\boldsymbol{F}(\mathbf{X}^{(j)},\boldsymbol{\xi}^{(j)})\right\|_F^2$$

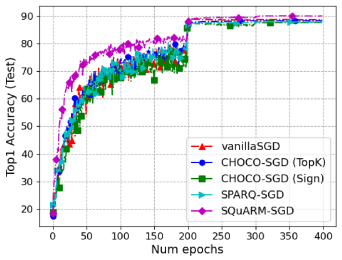


(a) For vanilla SGD

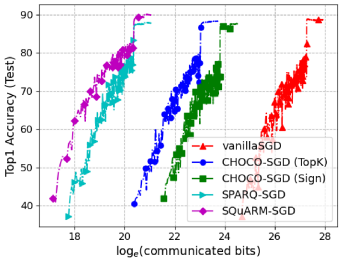


(b) For CHOCO-SGD ($TopK$)

Fig. 1 Increase in test accuracy when using momentum updates.



(a) Comparison of test accuracy



(b) Test accuracy vs. no. of bits

Fig. 2 Test performance comparison of SQuARM-SGD with other techniques.

From Proposition 3 (from page 10) to bound the stochastic gradient in the RHS of above equation gives us:

$$\sum_{t=0}^{T-1}\Lambda^{(t')} \leq \sum_{t=0}^{T-1}\sum_{j=0}^{t}\beta^{t-j}\left[2(M^2+1)(L^2\Xi^{(j)}+nG^2)\right]$$
$$+\sum_{t=0}^{T-1}\sum_{j=0}^{t}\beta^{t-j}\left[2(M^2+1)nB^2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(j)})\right\|_2^2+n\sigma^2\right]$$
$$\leq \frac{2(M^2+1)nG^2+n\sigma^2}{(1-\beta)}T+\frac{2(M^2+1)L^2}{(1-\beta)}\sum_{t=0}^{T-1}\Xi^{(t)}$$
$$+\frac{2(M^2+1)nB^2}{(1-\beta)}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|_2^2$$

Substituting the above bound in (49), we have:

$$\sum_{t=0}^{T-1}S^{(t)}\leq\eta^2T\left(\frac{8A\eta^2}{\alpha}+\left(\frac{32DH}{\alpha}\right)\left(\frac{2(M^2+1)nG^2+n\sigma^2}{(1-\beta)}\right)\right)$$
$$+\eta^2\left(\frac{32CH}{\alpha}+\left(\frac{32DH}{\alpha}\right)\frac{2(M^2+1)nB^2}{(1-\beta)}\right)\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(t)})\|^2$$
$$+\frac{64DH(M^2+1)L^2\eta^2}{\alpha(1-\beta)}\sum_{t=0}^{T-1}\Xi^{(t)}$$

Choose $\eta\leq\sqrt{\frac{\alpha(1-\beta)}{128DH(M^2+1)L^2}}$ and using that fact that $\Xi^{(t)}\leq S^{(t)}$ for all $t\in[T]$ and rearranging the summation term gives:

$$\frac{1}{T}\sum_{t=0}^{T-1}S^{(t)}\leq 2\eta^2J_1+2\eta^2J_2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|^2, \quad (50)$$

where $J_1=\left(\frac{8A\eta^2}{\alpha}+\left(\frac{32DH}{\alpha}\right)\left(\frac{2(M^2+1)nG^2+n\sigma^2}{(1-\beta)}\right)\right)$ and $J_2=\left(\frac{32CH}{\alpha}+\left(\frac{32DH}{\alpha}\right)\frac{2(M^2+1)nB^2}{(1-\beta)}\right)$.

VI. Experiments

In this section, we provide comparison of our proposed algorithm SQuARM-SGD, which uses momentum updates to CHOCO-SGD [20] and SPARQ-SGD [19] which consider compressed decentralized training (and local SGD, triggered communication for [19]) but do not incorporate momentum in their algorithms. We empirically demonstrate that using momentum based updates can increase the test performance of the learned model in large-scale decentralized training. We provide additional comparison experiments in Appendix J.

Setup. We match the setting in CHOCO-SGD, SPARQ-SGD and train ResNet20 [45] models on the CIFAR-10 [24] dataset with $n=8$ nodes connected in a ring topology. Learning rate follows a schedule: initialized to 0.2, warmup period of 5 epochs and has a decay of 10 at epoch 200 and 300; we stop training at epoch 400. For SQuARM-SGD, we use Nesterov momentum with a factor of $\beta=0.9$ and mini-batch size of 256. For either SPARQ-SGD [19] or CHOCO-SGD [20], we do not use momentum.[8] Matching [19], SQuARM-SGD consists of $H=5$ local iterations and we take top 1% elements of each tensor and only transmit the sign and norm of the result. The triggering threshold follows a schedule piecewise constant: initialized to 2.5 and increases by 1.5 after every 20 epochs till 350 epochs are complete, while maintaining that $c_t<1/\eta$ for all t. We compare performance of SQuARM-SGD against SPARQ-SGD (which uses $SignTopK$ compression, local iterations and threshold based communication), CHOCO-SGD with $Sign$, $TopK$ compression (taking top 1% of elements of the tensor) and decentralized vanilla SGD [17].

Results. We first demonstrate that performing momentum updates can lead to better test performance when training large scale machine learning models. Figure 1a and Figure 1b show test accuracy with and without momentum for vanilla SGD decentralized training and CHOCO-SGD (with $TopK$ compression), respectively. We observe that training with momentum updates improves test performance by 2-3%. Figure 2 compares the test performance for difference schemes, where SQuARM-SGD incorporates momentum updates (also theoretically analyzed) while CHOCO-SGD ($Sign$ or $TopK$ compression) and SPARQ-SGD ($SignTopK$ compression and local iterations) do not. Figure 2a shows that SQuARM-SGD has a better test performance than other methods by around 2% owing to momentum updates. Moreover, SQuARM-SGD reaches a higher test accuracy in relatively fewer epochs due to speedup by momentum. As SQuARM uses $SignTopK$ compression along with local iterations and triggering, it also achieves the target test accuracy of about 90% using significantly less communication bits[9] than either CHOCO-

[8] We note that while experimental results in [19], [20] were provided with momentum, they do not consider momentum in their analysis. Thus for a fair comparison, we consider our algorithm SQuARM-SGD with momentum updates while SPARQ-SGD, CHOCO-SGD are evaluated without momentum.

[9] As SPARQ-SGD [19] also uses $SignTopK$ compression with local iterations and event-triggering, it uses the same amount of communication bits as SQuARM-SGD although with an inferior test performance due to absence of momentum updates.

SGD or vanilla SGD training as demonstrated in Figure 2b.

REFERENCES

[1] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[2] N. Strom, "Scalable distributed DNN training using commodity GPU cloud computing," in *Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015, pp. 1488–1492.

[3] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," in *Proceedings of Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2017, pp. 440–445.

[4] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," in *International Conference on Learning Representations, ICLR*, 2018.

[5] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified SGD with Memory," in *Advances in Neural Information Processing Systems, NeurIPS*, 2018, pp. 4447–4458.

[6] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, "The convergence of sparsified gradient methods," in *Advances in Neural Information Processing Systems, NeurIPS*, 2018, pp. 5973–5983.

[7] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: communication-efficient SGD via gradient quantization and encoding," in *Advances in Neural Information Processing Systems, NIPS*, 2017, pp. 1709–1720.

[8] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "Terngrad: Ternary gradients to reduce communication in distributed deep learning," in *Advances in Neural Information Processing Systems, NIPS*, 2017, pp. 1508–1518.

[9] A. T. Suresh, F. X. Yu, S. Kumar, and H. B. McMahan, "Distributed mean estimation with limited communication," in *International Conference on Machine Learning, ICML*, 2017, pp. 3329–3337.

[10] S. P. Karimireddy, Q. Rebjock, S. U. Stich, and M. Jaggi, "Error feedback fixes signsgd and other gradient compression schemes," in *International Conference on Machine Learning, ICML*, 2019, pp. 3252–3261.

[11] J. Bernstein, Y.-X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," in *International Conference on Machine Learning, ICML*, 2018, pp. 560–569.

[12] D. Basu, D. Data, C. Karakus, and S. N. Diggavi, "Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations," in *Advances in Neural Information Processing Systems, NeurIPS*, 2019, pp. 14 668–14 679.

[13] S. U. Stich, "Local SGD Converges Fast and Communicates Little," in *International Conference on Learning Representations, ICLR*, 2019.

[14] H. Yu, S. Yang, and S. Zhu, "Parallel restarted SGD with faster convergence and less communication:demystifying why model averaging works for deep learning," in *AAAI Conference on Artificial Intelligence, AAAI*, 2019, pp. 5693–5700.

[15] G. F. Coppola, "Iterative parameter mixing for distributed large-margin training of structured predictors for natural language processing," Ph.D. dissertation, University of Edinburgh, UK, 2015.

[16] Y. Yan, T. Yang, Z. Li, Q. Lin, and Y. Yang, "A unified analysis of stochastic momentum methods for deep learning," in *Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI*, 2018, pp. 2955–2961.

[17] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," in *Advances in Neural Information Processing Systems, NIPS*, 2017, pp. 5330–5340.

[18] H. Tang, S. Gan, C. Zhang, T. Zhang, and J. Liu, "Communication compression for decentralized training," in *Advances in Neural Information Processing Systems, NeurIPS*, 2018, pp. 7663–7673.

[19] N. Singh, D. Data, J. George, and S. Diggavi, "SPARQ-SGD: Event-triggered and compressed communication in decentralized optimization," in *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE, 2020, pp. 3449–3456.

[20] A. Koloskova, T. Lin, S. U. Stich, and M. Jaggi, "Decentralized Deep Learning with Arbitrary Communication Compression," in *International Conference on Learning Representations, ICLR*, 2020.

[21] A. Koloskova, S. U. Stich, and M. Jaggi, "Decentralized Stochastic Optimization and Gossip Algorithms with Compressed Communication," in *International Conference on Machine Learning, ICML*, 2019, pp. 3478–3487.

[22] H. Tang, C. Yu, X. Lian, T. Zhang, and J. Liu, "Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression," in *International Conference on Machine Learning, ICML*, 2019, pp. 6155–6165.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016, pp. 770–778.

[24] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-10," *Canadian Institute for Advanced Research*, 2009.

[25] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, "Quantized decentralized consensus optimization," in *IEEE Conference on Decision and Control, CDC*, 2018, pp. 5838–5843.

[26] M. Assran, N. Loizou, N. Ballas, and M. Rabbat, "Stochastic gradient push for distributed deep learning," in *International Conference on Machine Learning, ICML*, 2019, pp. 344–353.

[27] T. Tatarenko and B. Touri, "Non-convex distributed optimization," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3744–3757, 2017.

[28] H. Yu, R. Jin, and S. Yang, "On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization," in *International Conference on Machine Learning, ICML*, 2019, pp. 7184–7193.

[29] J. Wang and G. Joshi, "Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms," *arXiv preprint arXiv:1808.07576*, 2018.

[30] J. Wang, A. K. Sahu, Z. Yang, G. Joshi, and S. Kar, "Matcha: Speeding up decentralized sgd via matching decomposition sampling," in *2019 Sixth Indian Control Conference (ICC)*. IEEE, 2019, pp. 299–300.

[31] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, "The marginal value of adaptive gradient methods in machine learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4151–4161.

[32] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. U. Stich, "A unified theory of decentralized SGD with changing topology and local updates," in *International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 5381–5393.

[33] J. Wang, V. Tantia, N. Ballas, and M. Rabbat, "SlowMo: Improving communication-efficient distributed sgd with slow momentum," in *International Conference on Learning Representations, ICLR*, 2020.

[34] S. Zheng, Z. Huang, and J. Kwok, "Communication-efficient distributed blockwise momentum sgd with error-feedback," in *Advances in Neural Information Processing Systems, NeurIPS*, 2019, pp. 11 446–11 456.

[35] W. P. M. H. Heemels, K. H. Johansson, and P. Tabuada, "An introduction to event-triggered and self-triggered control," in *IEEE Conference on Decision and Control, CDC*, 2012, pp. 3270–3285.

[36] D. V. Dimarogonas, E. Frazzoli, and K. H. Johansson, "Distributed event-triggered control for multi-agent systems," *IEEE Transactions on Automatic Control*, vol. 57, no. 5, pp. 1291–1297, 2012.

[37] G. S. Seyboth, D. V. Dimarogonas, and K. H. Johansson, "Event-based broadcasting for multi-agent average consensus," *Automatica*, vol. 49, no. 1, pp. 245–252, 2013.

[38] A. Girard, "Dynamic triggering mechanisms for event-triggered control," *IEEE Transactions on Automatic Control*, vol. 60, no. 7, pp. 1992–1997, 2015.

[39] Y. Liu, C. Nowzari, Z. Tian, and Q. Ling, "Asynchronous periodic event-triggered coordination of multi-agent systems," in *IEEE Conference on Decision and Control, CDC*, 2017, pp. 6696–6701.

[40] S. S. Kia, J. Cortés, and S. Martínez, "Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication," *Automatica*, vol. 55, pp. 254–264, 2015.

[41] W. Chen and W. Ren, "Event-triggered zero-gradient-sum distributed consensus optimization over directed networks," *Automatica*, vol. 65, pp. 90–97, 2016.

[42] W. Du, X. Yi, J. George, K. H. Johansson, and T. Yang, "Distributed optimization with dynamic event-triggered mechanisms," in *IEEE Conference on Decision and Control, CDC*, 2018, pp. 969–974.

[43] T. Chen, G. Giannakis, T. Sun, and W. Yin, "Lag: Lazily aggregated gradient for communication-efficient distributed learning," in *Advances in Neural Information Processing Systems, NeurIPS*, 2018, pp. 5050–5060.

[44] X. Lian, W. Zhang, C. Zhang, and J. Liu, "Asynchronous decentralized parallel stochastic gradient descent," in *International Conference on Machine Learning, ICML*, 2017, pp. 3043–3052.

[45] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, "Learning structured sparsity in deep neural networks," in *Advances in Neural Information Processing Systems, NIPS*, 2016, pp. 2074–2082.

APPENDIX A
PRELIMINARIES

Notation. Unless specified otherwise, for a vector $\mathbf{u}$, we write $\|\mathbf{u}\|$ to denote the ℓ_2-norm $\|\mathbf{u}\|_2$.

A. Vector and matrix inequalities

Fact 1. *Let $\mathbf{M} \in \mathbb{R}^{p\times q}$ be a matrix with entries $[m_{ij}]$, $i \in [p], j \in [q]$. The Frobenius norm of $\mathbf{M}$ is given by :*

$$\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{q} |m_{ij}|^2}$$

Consider any two matrices $\mathbf{A} \in \mathbb{R}^{d\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times n}$. Then the following holds:

$$\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_2 \tag{51}$$

Fact 2. *For any set of n vectors $\{\mathbf{a_1}, \ldots, \mathbf{a_n}\}$ where $\mathbf{a_i} \in \mathbb{R}^d$, we have:*

$$\left\|\sum_{i=1}^{n} \mathbf{a_i}\right\|^2 \leq n \sum_{i=1}^{n} \|\mathbf{a_i}\|^2 \tag{52}$$

Fact 3. *For any two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, for all $\gamma > 0$, we have:*

$$2\langle \mathbf{a}, \mathbf{b}\rangle \leq \gamma \|\mathbf{a}\|^2 + \gamma^{-1} \|\mathbf{b}\|^2 \tag{53}$$

Fact 4. *For any two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, for all $\alpha > 0$, we have:*

$$\|\mathbf{a}+\mathbf{b}\|^2 \leq (1+\alpha)\|\mathbf{a}\|^2 + (1+\alpha^{-1})\|\mathbf{b}\|^2 \tag{54}$$

Similar inequality holds for matrices in Frobenius norm, i.e., for any two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times q}$ and for any $\alpha > 0$, we have

$$\|\mathbf{A}+\mathbf{B}\|_F^2 \leq (1+\alpha)\|\mathbf{A}\|_F^2 + (1+\alpha^{-1})\|\mathbf{B}\|_F^2$$

B. Properties of functions

Definition 2 (Smoothness)**.** *A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth with parameter $L \geq 0$ if*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \tag{55}$$

Lemma 9. *Let f be an L-smooth function with global minimizer $\mathbf{x}^*$. We have*

$$\|\nabla f(\mathbf{x})\|^2 \leq 2L(f(\mathbf{x}) - f(\mathbf{x}^*)). \tag{56}$$

Proof. By definition of L-smoothness, we have

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2.$$

Taking infimum over y yields:

$$\begin{aligned}
\inf_{\mathbf{y}} f(\mathbf{y}) &\leq \inf_{\mathbf{y}} \left(f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2 \right) \\
&\overset{(a)}{=} \inf_{\mathbf{v}:\|\mathbf{v}\|=1} \inf_{t} \left(f(\mathbf{x}) + t\langle \nabla f(\mathbf{x}), \mathbf{v}\rangle + \frac{Lt^2}{2} \right) \\
&\overset{(b)}{=} \inf_{\mathbf{v}:\|\mathbf{v}\|=1} \left(f(\mathbf{x}) - \frac{1}{2L}\langle \nabla f(\mathbf{x}), \mathbf{v}\rangle^2 \right) \\
&\overset{(c)}{=} \left(f(\mathbf{x}) - \frac{1}{2L}\|\nabla f(\mathbf{x})\|^2 \right)
\end{aligned}$$

The value of t that minimizes the RHS of (a) is $t = -\frac{1}{L}\langle \nabla f(\mathbf{x}), \mathbf{v}\rangle$, this implies (b); (c) follows from the Cauchy-Schwartz inequality: $\langle \mathbf{u}, \mathbf{v}\rangle \leq \|\mathbf{u}\|\|\mathbf{v}\|$, where equality is achieved whenever $u = v$. Now, substituting $\inf_{\mathbf{y}} f(\mathbf{y}) = f(\mathbf{x}^*)$ in the RHS of (c) yields the result. □

APPENDIX B
PRELIMINARIES FOR CONVERGENCE WITH RELAXED ASSUMPTIONS

Proof of Proposition 1. This simply follows from the independence of the randomness used in sampling stochastic gradients at different workers. □

Proof of Proposition 2. We want to show the following bound on $\mathbb{E}\left\|\mathbf{V}^{(t)}\right\|_F^2$ for any t:

$$\mathbb{E}\left\|\mathbf{V}^{(t)}\right\|_F^2 \leq \frac{1}{(1-\beta)}\sum_{k=0}^{t}\beta^{t-k}\mathbb{E}\left\|\nabla \boldsymbol{F}(\mathbf{X}^{(k)},\boldsymbol{\xi}^{(k)})\right\|_F^2.$$

For any t, let $\theta_t = \sum_{k=0}^{t}\beta^{t-k}$.

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{V}^{(t)}\right\|_F^2 &= \mathbb{E}\left\|\sum_{k=0}^{t}\beta^{t-k}\nabla \boldsymbol{F}(\mathbf{X}^{(k)},\boldsymbol{\xi}^{(k)})\right\|_F^2 \\
&= \theta_t^2\mathbb{E}\left\|\sum_{k=0}^{t}\frac{\beta^{t-k}}{\theta_t}\nabla \boldsymbol{F}(\mathbf{X}^{(k)},\boldsymbol{\xi}^{(k)})\right\|_F^2 \\
&\leq \theta_t\sum_{k=0}^{t}\beta^{t-k}\mathbb{E}\left\|\nabla \boldsymbol{F}(\mathbf{X}^{(k)},\boldsymbol{\xi}^{(k)})\right\|_F^2 \\
&\leq \frac{1}{1-\beta}\sum_{k=0}^{t}\beta^{t-k}\mathbb{E}\left\|\nabla \boldsymbol{F}(\mathbf{X}^{(k)},\boldsymbol{\xi}^{(k)})\right\|_F^2. \qquad (57)
\end{aligned}$$

□

APPENDIX C
OMITTED DETAILS FROM SECTION V

A. Omitted Details from Section V-C

Lemma 10. *We have the following bounds on P_1 and P_2 (which are defined in* (16)*):*

$$\begin{aligned}
P_1 &\leq -\frac{\eta}{2(1-\beta)}\left\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\right\|^2 + \frac{\eta L^2}{2n(1-\beta)}\sum_{i=1}^{n}\left\|\widetilde{\mathbf{x}}^{(t)}-\mathbf{x}_i^{(t)}\right\|^2, \\
P_2 &\leq \frac{\sigma^2}{n} + \frac{2(M^2+n)L^2}{n^2}\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)}-\widetilde{\mathbf{x}}^{(t)}\right\|_2^2 + \frac{2(M^2+n)}{n}\big(G^2 + B^2\left\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\right\|_2^2\big).
\end{aligned}$$

Proof.

$$\begin{aligned}
P_1 &= -\left\langle \nabla f(\widetilde{\mathbf{x}}^{(t)}), \frac{\eta}{(1-\beta)}\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})\right\rangle \\
&= -\left\langle \nabla f(\widetilde{\mathbf{x}}^{(t)}), \frac{\eta}{(1-\beta)}\frac{1}{n}\sum_{i=1}^{n}\Big(\nabla f_i(\mathbf{x}_i^{(t)}) - \nabla f_i(\widetilde{\mathbf{x}}^{(t)}) + \nabla f_i(\widetilde{\mathbf{x}}^{(t)})\Big)\right\rangle \\
&= -\left\langle \nabla f(\widetilde{\mathbf{x}}^{(t)}), \frac{\eta}{(1-\beta)}\nabla f(\widetilde{\mathbf{x}}^{(t)})\right\rangle + \frac{\eta}{(1-\beta)}\frac{1}{n}\sum_{i=1}^{n}\left\langle \nabla f(\widetilde{\mathbf{x}}^{(t)}), \nabla f_i(\widetilde{\mathbf{x}}^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})\right\rangle \\
&\overset{(b)}{\leq} -\frac{\eta}{(1-\beta)}\left\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\right\|^2 + \frac{\eta}{2(1-\beta)}\left\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\right\|^2 + \frac{\eta}{2(1-\beta)}\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(\widetilde{\mathbf{x}}^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 \\
&\overset{(c)}{\leq} -\frac{\eta}{2(1-\beta)}\left\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\right\|^2 + \frac{\eta L^2}{2n(1-\beta)}\sum_{i=1}^{n}\left\|\widetilde{\mathbf{x}}^{(t)}-\mathbf{x}_i^{(t)}\right\|^2,
\end{aligned}$$

where (b) follows from $\langle \mathbf{a},\mathbf{b}\rangle \leq \frac{1}{2}(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2)$ and (c) follows from the L-smoothness of f_i.

For bounding P_2, we will use Proposition 1.

$$\begin{aligned}
P_2 &= \mathbb{E}_{\xi_{(t)}}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)})\right\|^2 \\
&\overset{(d)}{=} \mathbb{E}_{\xi_{(t)}}\left\|\frac{1}{n}\sum_{i=1}^{n}\big(\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})\big)\right\|^2 + \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(e)}}{\leq} \frac{\sigma^2}{n} + \frac{M^2}{n^2}\sum_{i=1}^{n}\left\|\nabla f_i(\mathbf{x}_i^{(t)})\right\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 \\
&= \frac{\sigma^2}{n} + \frac{(M^2+n)}{n^2}\sum_{i=1}^{n}\left\|\nabla f_i(\mathbf{x}_i^{(t)})\right\|_2^2 \\
&\leq \frac{\sigma^2}{n} + \frac{2(M^2+n)}{n^2}\sum_{i=1}^{n}\left\|\nabla f_i(\mathbf{x}_i^{(t)}) - \nabla f_i(\widetilde{\mathbf{x}}^{(t)})\right\|_2^2 + \frac{2(M^2+n)}{n^2}\sum_{i=1}^{n}\left\|\nabla f_i(\widetilde{\mathbf{x}}^{(t)})\right\|_2^2 \\
&\overset{\text{(f)}}{\leq} \frac{\sigma^2}{n} + \frac{2(M^2+n)L^2}{n^2}\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)} - \widetilde{\mathbf{x}}^{(t)}\right\|_2^2 + \frac{2(M^2+n)}{n}\big(G^2 + B^2\left\|\nabla f(\widetilde{\mathbf{x}}^{(t)})\right\|_2^2\big)
\end{aligned} \tag{58}
$$

Here, (d) follows because the randomness used for sampling the unbiased stochastic gradients across workers is independent of each other, (e) follows from (11), and (f) follows from the L-smoothness of f_i and (4). □

Lemma (Restating Lemma 5)**.** *Consider the deviation of the global average parameter $\overline{\mathbf{x}}^{(t)}$ and the virtual sequence $\widetilde{\mathbf{x}}^{(t)}$ defined in* (13) *for constant stepsize η. Then at any time step t, the following holds:*

$$
\left\|\overline{\mathbf{x}}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\right\|^2 \leq \frac{\beta^4\eta^2}{(1-\beta)^3}\sum_{\tau=0}^{t-1}\left[\beta^{t-\tau-1}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(\tau)}, \xi_i^{(\tau)})\right\|^2\right] \tag{59}
$$

Proof. Using the definition of $\tilde{\mathbf{x}}^{(t)}$ as in (13), we have:

$$
\left\|\bar{\mathbf{x}}^{(t)} - \tilde{\mathbf{x}}^{(t)}\right\|^2 = \left\|\bar{\mathbf{x}}^{(t)} - \tilde{\mathbf{x}}^{(t)}\right\|^2 = \frac{\beta^4\eta^2}{(1-\beta)^2}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i^{(t-1)}\right\|^2
$$

Define $\theta_{t-1} = \sum_{k=0}^{t-1}\beta^{1-t-k} = \frac{1-\beta^t}{1-\beta}$. Thus we can expand the term in the norm as:

$$
\begin{aligned}
&= \frac{\beta^4\eta^2}{(1-\beta)^2}\theta_{t-1}^2\left\|\sum_{k=0}^{t-1}\frac{\beta^{t-1-k}}{\theta_{t-1}}\frac{1}{n}\sum_{i=1}^{n}\nabla F(\mathbf{x}_i^{(k)}, \xi_i^{(k)})\right\|^2 \\
&\leq \frac{\beta^4\eta^2}{(1-\beta)^2}\theta_{t-1}^2\sum_{k=0}^{t-1}\frac{\beta^{t-1-k}}{\theta_{t-1}}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F(\mathbf{x}_i^{(k)}, \xi_i^{(k)})\right\|^2 \\
&= \frac{\beta^4\eta^2}{(1-\beta)^2}\theta_{t-1}\sum_{k=0}^{t-1}\beta^{t-1-k}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F(\mathbf{x}_i^{(k)}, \xi_i^{(k)})\right\|^2 \\
&\leq \frac{\beta^4\eta^2}{(1-\beta)^3}\sum_{\tau=0}^{t-1}\left[\beta^{t-\tau-1}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(\tau)}, \xi_i^{(\tau)})\right\|^2\right]
\end{aligned}
$$

Where the first inequality follows from Jensen's inequality and the second inequality follows from noting that $\theta_t \leq \frac{1}{1-\beta}$. This completes the proof. □

Proof of Lemma 6. We have already bounded the expectation term in (18) – the same bound holds when expectation is taken w.r.t. the entire past. Substituting that bound – i.e., $\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(\tau)}, \xi_i^{(\tau)})\right\|^2 \leq \frac{\sigma^2}{n} + \frac{(M^2+n)}{n^2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{(\tau)})\right\|_2^2$ – from (58) into (23) gives

$$
\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\left[\beta^{t-\tau-1}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(\tau)}, \xi_i^{(\tau)})\right\|^2\right] &\leq \frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\beta^{t-\tau-1}\frac{\sigma^2}{n} \\
&+ \frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\beta^{t-\tau-1}\frac{(M^2+n)}{n^2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{(\tau)})\right\|_2^2
\end{aligned} \tag{60}
$$

Now we bound both the terms of (60) separately.

$$
\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\beta^{t-\tau-1}\frac{\sigma^2}{n} = \frac{\sigma^2}{n}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\beta^{t-\tau-1} \leq \frac{\sigma^2}{n(1-\beta)}. \tag{61}
$$

$$
\frac{1}{T}\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\beta^{t-\tau-1}\frac{(M^2+n)}{n^2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{(\tau)})\right\|_2^2 = \frac{1}{T}\sum_{\tau=0}^{T-2}\sum_{t=\tau+1}^{T-1}\beta^{t-\tau-1}\frac{(M^2+n)}{n^2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{(\tau)})\right\|_2^2
$$

$$
\begin{aligned}
&= \frac{(M^2+n)}{n^2}\frac{1}{T}\sum_{\tau=0}^{T-2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{(\tau)})\right\|_2^2\sum_{t=\tau+1}^{T-1}\beta^{t-\tau-1}\\
&\leq \frac{(M^2+n)}{n^2(1-\beta)}\frac{1}{T}\sum_{\tau=0}^{T-2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{(\tau)})\right\|_2^2\\
&\leq \frac{2(M^2+n)}{n^2(1-\beta)}\frac{1}{T}\sum_{\tau=0}^{T-2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\mathbf{x}_i^{(\tau)})-\nabla f_i(\overline{\mathbf{x}}^{(\tau)})\right\|_2^2+\frac{2(M^2+n)}{n^2(1-\beta)}\frac{1}{T}\sum_{\tau=0}^{T-2}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_i(\overline{\mathbf{x}}^{(\tau)})\right\|_2^2\\
&\leq \frac{2(M^2+n)}{n^2(1-\beta)}\frac{1}{T}\sum_{\tau=0}^{T-2}\sum_{i=1}^{n}L^2\mathbb{E}\left\|\mathbf{x}_i^{(\tau)}-\overline{\mathbf{x}}^{(\tau)}\right\|_2^2+\frac{2(M^2+n)}{n(1-\beta)}\frac{1}{T}\sum_{\tau=0}^{T-2}\left(G^2+B^2\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(\tau)})\|_2^2\right)\\
&\leq \frac{2(M^2+n)L^2}{n^2(1-\beta)}\frac{1}{T}\sum_{\tau=0}^{T-2}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}_i^{(\tau)}-\overline{\mathbf{x}}^{(\tau)}\right\|_2^2+\frac{2(M^2+n)G^2}{n(1-\beta)}+\frac{2(M^2+n)B^2}{n(1-\beta)}\frac{1}{T}\sum_{\tau=0}^{T-2}\mathbb{E}\|\nabla f(\overline{\mathbf{x}}^{(\tau)})\|_2^2 \qquad (62)
\end{aligned}
$$

Substituting the bounds from (61), (62) into (60) yields (23), which proves Lemma 6. □

B. Omitted Details from Section V-D

1) Proof of Lemma 7: In this section we will prove Lemma 7.

Proof. We show the following bound in Lemma 11 (provided at the end of this section):

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{X}^{((m+1)H)}-\overline{\mathbf{X}}^{((m+1)H)}\right\|_F^2 \leq{}& \vartheta_1\mathbb{E}\left\|\mathbf{X}^{(mH)}-\overline{\mathbf{X}}^{(mH)}\right\|_F^2+\vartheta_2\mathbb{E}\left\|\mathbf{X}^{(mH)}-\widehat{\mathbf{X}}^{((m+1)H)}\right\|_F^2\\
&+\vartheta_3\eta^2\mathbb{E}\left\|\sum_{t'=mH}^{(m+1)H-1}\beta\mathbf{V}^{(t')}+\nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right\|_F^2, \qquad (63)
\end{aligned}
$$

where $\vartheta_1=(1+\alpha_5^{-1})R_1$, $\vartheta_2=(1+\alpha_5^{-1})R_2$, and $\vartheta_3=(R_1+R_2)(1+\alpha_5)$.

We want to write the second expectation term $\mathbb{E}\left\|\mathbf{X}^{(mH)}-\widehat{\mathbf{X}}^{((m+1)H)}\right\|_F^2$ on the RHS of (63) in terms of $\mathbb{E}\left\|\mathbf{X}^{(mH)}-\widehat{\mathbf{X}}^{(mH)}\right\|_F^2$. For that, first we define

$$
\mathbf{X}^{((m+1/2)H)}:=\mathbf{X}^{(mH)}-\eta\sum_{t'=mH}^{(m+1)H-1}\left(\beta\mathbf{V}^{(t')}+\nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right). \qquad (64)
$$

$$
\begin{aligned}
\mathbb{E}\left\|\mathbf{X}^{(mH)}-\widehat{\mathbf{X}}^{((m+1)H)}\right\|_F^2 &= \mathbb{E}\left\|\mathbf{X}^{(mH)}-\left(\widehat{\mathbf{X}}^{(mH)}+\mathcal{C}\left(\mathbf{X}^{((m+1/2)H)}-\widehat{\mathbf{X}}^{(mH)}\right)\right)\right\|_F^2\\
&= \mathbb{E}\left\|\mathbf{X}^{((m+1/2)H)}-\widehat{\mathbf{X}}^{(mH)}-\mathcal{C}\left(\mathbf{X}^{((m+1/2)H)}-\widehat{\mathbf{X}}^{(mH)}\right)+\mathbf{X}^{(mH)}-\mathbf{X}^{((m+1/2)H)}\right\|_F^2\\
&\leq (1+\tau_1)(1-\omega)\mathbb{E}\left\|\mathbf{X}^{((m+1/2)H)}-\widehat{\mathbf{X}}^{(mH)}\right\|_F^2+(1+\tau_1^{-1})\mathbb{E}\left\|\mathbf{X}^{(mH)}-\mathbf{X}^{((m+1/2)H)}\right\|_F^2\\
&= (1+\tau_1)(1-\omega)\mathbb{E}\left\|\mathbf{X}^{((m+1/2)H)}-\mathbf{X}^{(mH)}+\mathbf{X}^{(mH)}-\widehat{\mathbf{X}}^{(mH)}\right\|_F^2\\
&\qquad+(1+\tau_1^{-1})\mathbb{E}\left\|\mathbf{X}^{(mH)}-\mathbf{X}^{((m+1/2)H)}\right\|_F^2\\
&\leq (1+\tau_1)(1-\omega)(1+\tau_2)\mathbb{E}\left\|\mathbf{X}^{(mH)}-\widehat{\mathbf{X}}^{(mH)}\right\|_F^2\\
&\qquad+\left((1+\tau_1^{-1})+(1+\tau_1)(1-\omega)(1+\tau_2^{-1})\right)\mathbb{E}\left\|\mathbf{X}^{(mH)}-\mathbf{X}^{((m+1/2)H)}\right\|_F^2\\
&\leq \chi_1\mathbb{E}\left\|\mathbf{X}^{(mH)}-\widehat{\mathbf{X}}^{(mH)}\right\|_F^2+\chi_2\eta^2\mathbb{E}\left\|\sum_{t'=mH}^{(m+1)H-1}\left(\beta\mathbf{V}^{(t')}+\nabla\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right)\right\|_F^2, \qquad (65)
\end{aligned}
$$

where $\chi_1=(1+\tau_1)(1-\omega)(1+\tau_2)$ and $\chi_2=\left((1+\tau_1^{-1})+(1+\tau_1)(1-\omega)(1+\tau_2^{-1})\right)$.

Substituting this back in (63) yields (27), which proves Lemma 7. □

Lemma 11. *We have*

$$
\mathbb{E}\|\mathbf{X}^{((m+1)H)}-\bar{\mathbf{X}}^{((m+1)H)}\|_F^2\leq R_1(1+\alpha_5^{-1})\mathbb{E}\left\|\bar{\mathbf{X}}^{(mH)}-\mathbf{X}^{(mH)}\right\|^2+R_2(1+\alpha_5^{-1})\mathbb{E}\left\|\hat{\mathbf{X}}^{((m+1)H)}-\mathbf{X}^{(mH)}\right\|^2
$$

$$+ (1+\alpha_5)(R_1+R_2)\eta^2 \left\| \sum_{t'=(mH)}^{((m+1)H)-1} (\beta \mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})) \right\|_F^2$$

Proof. Using the update equations of $\mathbf{X}^{((m+1)H)}$ in matrix form given in (5)-(8) in Section IV, we have:

$$\|\mathbf{X}^{((m+1)H)} - \bar{\mathbf{X}}^{((m+1)H)}\|_F^2 = \|\mathbf{X}^{((m+1/2)H)} - \bar{\mathbf{X}}^{((m+1)H)} + \gamma \hat{\mathbf{X}}^{((m+1)H)}(\mathbf{W}-\mathbf{I})\|_F^2$$

Noting that $\bar{\mathbf{X}}^{((m+1)H)} = \bar{\mathbf{X}}^{((m+1/2)H)}$ (from (10)) and $\bar{\mathbf{X}}^{((m+1/2)H)}(\mathbf{W}-\mathbf{I}) = 0$ (from (9)), we get:

$$\|\mathbf{X}^{((m+1)H)} - \bar{\mathbf{X}}^{((m+1)H)}\|_F^2 = \|(\mathbf{X}^{((m+1/2)H)} - \bar{\mathbf{X}}^{((m+1/2)H)})((1-\gamma)\mathbf{I} + \gamma\mathbf{W}) + \gamma(\hat{\mathbf{X}}^{((m+1)H)} - \mathbf{X}^{((m+1/2)H)})(\mathbf{W}-\mathbf{I})\|_F^2$$

For any positive constant[10] α_1, we have:

$$\|\mathbf{X}^{((m+1)H)} - \bar{\mathbf{X}}^{((m+1)H)}\|_F^2 \leq (1+\alpha_1)\|(\mathbf{X}^{((m+1/2)H)} - \bar{\mathbf{X}}^{((m+1/2)H)})((1-\gamma)\mathbf{I} + \gamma\mathbf{W})\|_F^2 + (1+\alpha_1^{-1})\|\gamma(\hat{\mathbf{X}}^{((m+1)H)} - \mathbf{X}^{((m+1/2)H)})(\mathbf{W}-\mathbf{I})\|_F^2$$

Using $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_2$ for any matrices $\mathbf{A}, \mathbf{B}$, we have:

$$\|\mathbf{X}^{((m+1)H)} - \bar{\mathbf{X}}^{((m+1)H)}\|_F^2 \leq (1+\alpha_1)\|(\mathbf{X}^{((m+1/2)H)} - \bar{\mathbf{X}}^{((m+1/2)H)})((1-\gamma)\mathbf{I} + \gamma\mathbf{W})\|_F^2 + (1+\alpha_1^{-1})\gamma^2\|(\hat{\mathbf{X}}^{((m+1)H)} - \mathbf{X}^{((m+1/2)H)})\|_F^2 . \|(\mathbf{W}-\mathbf{I})\|_2^2 \tag{66}$$

To bound the first term in (150), we use the triangle inequality for Frobenius norm, giving us:

$$\|(\mathbf{X}^{((m+1/2)H)} - \bar{\mathbf{X}}^{((m+1/2)H)})((1-\gamma)\mathbf{I} + \gamma\mathbf{W})\|_F \leq (1-\gamma)\|\mathbf{X}^{((m+1/2)H)} - \bar{\mathbf{X}}^{((m+1/2)H)}\|_F + \gamma\|(\mathbf{X}^{((m+1/2)H)} - \bar{\mathbf{X}}^{((m+1/2)H)})\mathbf{W}\|_F$$

Since $\left(\mathbf{X}^{((m+1/2)H)} - \bar{\mathbf{X}}^{((m+1/2)H)}\right)\frac{\mathbf{1}\mathbf{1}^T}{n} = \mathbf{0}$ (from (9)), adding this inside the last term above, we get:

$$\|(\mathbf{X}^{((m+1/2)H)} - \bar{\mathbf{X}}^{((m+1/2)H)})((1-\gamma)\mathbf{I} + \gamma\mathbf{W})\|_F \leq (1-\gamma)\|\mathbf{X}^{((m+1/2)H)} - \bar{\mathbf{X}}^{((m+1/2)H)}\|_F + \gamma\left\|(\mathbf{X}^{((m+1/2)H)} - \bar{\mathbf{X}}^{((m+1/2)H)})\left(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\right\|_F$$

Using $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F\|\mathbf{B}\|_2$ and then using (112) from Fact 3 with $k=1$, we can simplify the above to:

$$\|(\mathbf{X}^{((m+1/2)H)} - \bar{\mathbf{X}}^{((m+1/2)H)})((1-\gamma)\mathbf{I} + \gamma\mathbf{W})\|_F \leq (1-\gamma\delta)\|\mathbf{X}^{((m+1/2)H)} - \bar{\mathbf{X}}^{((m+1/2)H)}\|_F$$

Substituting the above in (150) and using $\lambda = \max_i\{1-\lambda_i(\mathbf{W})\} \Rightarrow \|\mathbf{W}-\mathbf{I}\|_2^2 \leq \lambda^2$, we get:

$$\|\mathbf{X}^{((m+1)H)} - \bar{\mathbf{X}}^{((m+1)H)}\|_F^2 \leq (1+\alpha_1)(1-\gamma\delta)^2\|\mathbf{X}^{((m+1/2)H)} - \bar{\mathbf{X}}^{((m+1/2)H)}\|_F^2 + (1+\alpha_1^{-1})\gamma^2\lambda^2\|\mathbf{X}^{((m+1/2)H)} - \hat{\mathbf{X}}^{((m+1)H)}\|_F^2$$

Taking expectation w.r.t. the entire process, we have:

$$\mathbb{E}\|\mathbf{X}^{((m+1)H)} - \bar{\mathbf{X}}^{((m+1)H)}\|_F^2 \leq (1+\alpha_1)(1-\gamma\delta)^2\mathbb{E}\|\mathbf{X}^{((m+1/2)H)} - \bar{\mathbf{X}}^{((m+1/2)H)}\|_F^2 + (1+\alpha_1^{-1})\gamma^2\lambda^2\mathbb{E}\|\mathbf{X}^{((m+1/2)H)} - \hat{\mathbf{X}}^{((m+1)H)}\|_F^2$$

Define $R_1 = (1+\alpha_1)(1-\gamma\delta)^2, R_2 = (1+\alpha_1^{-1})\gamma^2\lambda^2$. Using the update steps of algorithm given in equations (6) and (10) (given in Section IV), we have:

$$\mathbb{E}\|\mathbf{X}^{((m+1)H)} - \bar{\mathbf{X}}^{((m+1)H)}\|_F^2 \leq R_1\mathbb{E}\left\|\bar{\mathbf{X}}^{(mH)} - \mathbf{X}^{(mH)} - \sum_{t'=mH}^{(m+1)H-1} \eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}))\left(\frac{\mathbf{1}\mathbf{1}^T}{n} - I\right)\right\|_F^2 + R_2\mathbb{E}\left\|\hat{\mathbf{X}}^{((m+1)H)} - \mathbf{X}^{(mH)} + \sum_{t'=mH}^{(m+1)H-1} \eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}))\right\|_F^2$$

[10]For any two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times q}$ and for any $\alpha > 0$, we have the following relationship for the Frobenius norm:

$$\|\mathbf{A}+\mathbf{B}\|_F^2 \leq (1+\alpha)\|\mathbf{A}\|_F^2 + (1+\alpha^{-1})\|\mathbf{B}\|_F^2$$

Thus, for any $\alpha_5 > 0$ (using Footnote 11), we have:

$$\begin{aligned}\mathbb{E}\|\mathbf{X}^{((m+1)H)} - \bar{\mathbf{X}}^{((m+1)H)}\|_F^2 \leq\ & R_1(1+\alpha_5^{-1})\mathbb{E}\left\|\bar{\mathbf{X}}^{(mH)} - \mathbf{X}^{(mH)}\right\|^2 + R_2(1+\alpha_5^{-1})\mathbb{E}\left\|\hat{\mathbf{X}}^{((m+1)H)} - \mathbf{X}^{(mH)}\right\|^2 \\ &+ R_1(1+\alpha_5)\mathbb{E}\left\|\sum_{t'=(mH)}^{((m+1)H)-1}\eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}))\left(\frac{\mathbf{1}\mathbf{1}^T}{n} - I\right)\right\|_F^2 \\ &+ R_2(1+\alpha_5)\mathbb{E}\left\|\sum_{t'=(mH)}^{((m+1)H)-1}\eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}))\right\|_F^2\end{aligned}$$

Using $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_2$ to split the third term, and then using the bound $\left\|\frac{\mathbf{1}\mathbf{1}^T}{n} - \mathbf{I}\right\|_2 = 1$ (which is shown in Claim 2 in Appendix D in supplementary), the above can be rewritten as:

$$\begin{aligned}\mathbb{E}\|\mathbf{X}^{((m+1)H)} - \bar{\mathbf{X}}^{((m+1)H)}\|_F^2 \leq\ & R_1(1+\alpha_5^{-1})\mathbb{E}\left\|\bar{\mathbf{X}}^{(mH)} - \mathbf{X}^{(mH)}\right\|^2 + R_2(1+\alpha_5^{-1})\mathbb{E}\left\|\hat{\mathbf{X}}^{((m+1)H)} - \mathbf{X}^{(mH)}\right\|^2 \\ &+ (1+\alpha_5)(R_1+R_2)\eta^2\left\|\sum_{t'=(mH)}^{((m+1)H)-1}(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}))\right\|_F^2\end{aligned}$$

□

2) Proof of Lemma 8: In this section, we prove Lemma 8.

Proof.

$$\begin{aligned}\mathbb{E}\left\|\mathbf{X}^{((m+1)H)} - \widehat{\mathbf{X}}^{((m+1)H)}\right\|_F^2 &= \mathbb{E}\left\|\mathbf{X}^{((m+1)H)} - \left(\widehat{\mathbf{X}}^{(mH)} + \mathcal{C}\left(\mathbf{X}^{((m+1/2)H)} - \widehat{\mathbf{X}}^{(mH)}\right)\right)\right\|_F^2 \\ &= \mathbb{E}\left\|\mathbf{X}^{((m+1/2)H)} - \widehat{\mathbf{X}}^{(mH)} - \mathcal{C}\left(\mathbf{X}^{((m+1/2)H)} - \widehat{\mathbf{X}}^{(mH)}\right) + \mathbf{X}^{((m+1)H)} - \mathbf{X}^{((m+1/2)H)}\right\|_F^2 \\ &\leq (1+\tau_3)(1-\omega)\underbrace{\mathbb{E}\left\|\mathbf{X}^{((m+1/2)H)} - \widehat{\mathbf{X}}^{(mH)}\right\|_F^2}_{=:\ T_1} + (1+\tau_3^{-1})\underbrace{\mathbb{E}\left\|\mathbf{X}^{((m+1)H)} - \mathbf{X}^{((m+1/2)H)}\right\|_F^2}_{=:\ T_2} \qquad (67)\end{aligned}$$

Now we bound T_1 and T_2.

$$\begin{aligned}T_1 &= \mathbb{E}\left\|\mathbf{X}^{((m+1/2)H)} - \widehat{\mathbf{X}}^{(mH)}\right\|_F^2 \\ &= \mathbb{E}\left\|\mathbf{X}^{(mH)} - \eta\sum_{t'=mH}^{(m+1)H-1}\left(\beta\mathbf{V}^{(t')} + \nabla\boldsymbol{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})\right) - \widehat{\mathbf{X}}^{(mH)}\right\|_F^2 \\ &\leq (1+\tau_4)\mathbb{E}\left\|\mathbf{X}^{(mH)} - \widehat{\mathbf{X}}^{(mH)}\right\|_F^2 + (1+\tau_4^{-1})\eta^2\mathbb{E}\left\|\sum_{t'=mH}^{(m+1)H-1}\left(\beta\mathbf{V}^{(t')} + \nabla\boldsymbol{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})\right)\right\|_F^2 \qquad (68)\end{aligned}$$

$$\begin{aligned}T_2 &= \mathbb{E}\left\|\mathbf{X}^{((m+1)H)} - \mathbf{X}^{((m+1/2)H)}\right\|_F^2 \\ &= \mathbb{E}\left\|\mathbf{X}^{((m+1/2)H)} + \gamma\widehat{\mathbf{X}}^{((m+1)H)}(\mathbf{W}-\mathbf{I}) - \mathbf{X}^{((m+1/2)H)}\right\|_F^2 \\ &= \gamma^2\mathbb{E}\left\|\widehat{\mathbf{X}}^{((m+1)H)}(\mathbf{W}-\mathbf{I})\right\|_F^2 \\ &= \gamma^2\mathbb{E}\left\|\left(\widehat{\mathbf{X}}^{((m+1)H)} - \overline{\mathbf{X}}^{((m+1/2)H)}\right)(\mathbf{W}-\mathbf{I})\right\|_F^2 && \text{(Since } \overline{\mathbf{X}}^{((m+1/2)H)}(\mathbf{W}-\mathbf{I}) = \mathbf{0}\text{)} \\ &\leq \gamma^2\lambda^2\mathbb{E}\left\|\widehat{\mathbf{X}}^{((m+1)H)} - \overline{\mathbf{X}}^{((m+1/2)H)}\right\|_F^2 && \text{(Since } \|\mathbf{W}-\mathbf{I}\|_2 = \lambda\text{)} \\ &= \gamma^2\lambda^2\mathbb{E}\left\|\widehat{\mathbf{X}}^{((m+1)H)} - \left(\overline{\mathbf{X}}^{(mH)} - \eta\sum_{t'=mH}^{(m+1)H-1}\left(\beta\mathbf{V}^{(t')} + \nabla\boldsymbol{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})\right)\right)\right\|_F^2 \\ &\leq \phi_1\underbrace{\mathbb{E}\left\|\widehat{\mathbf{X}}^{((m+1)H)} - \overline{\mathbf{X}}^{(mH)}\right\|_F^2}_{=:\ T_3} + \phi_2\eta^2\mathbb{E}\left\|\sum_{t'=mH}^{(m+1)H-1}\left(\beta\mathbf{V}^{(t')} + \nabla\boldsymbol{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})\right)\right\|_F^2, \qquad (69)\end{aligned}$$

where $\phi_1 = \gamma^2\lambda^2(1+\tau_5)$ and $\phi_2 = \gamma^2\lambda^2(1+\tau_5^{-1})$.

$$\begin{aligned}
T_3 &= \mathbb{E}\left\|\widehat{\mathbf{X}}^{((m+1)H)} - \overline{\mathbf{X}}^{(mH)}\right\|_F^2 \\
&= \mathbb{E}\left\|\widehat{\mathbf{X}}^{((m+1)H)} - \mathbf{X}^{(mH)} + \mathbf{X}^{(mH)} - \overline{\mathbf{X}}^{(mH)}\right\|_F^2 \\
&\leq (1+\tau_6)\mathbb{E}\left\|\mathbf{X}^{(mH)} - \overline{\mathbf{X}}^{(mH)}\right\|_F^2 + (1+\tau_6^{-1})\mathbb{E}\left\|\widehat{\mathbf{X}}^{((m+1)H)} - \mathbf{X}^{(mH)}\right\|_F^2 \\
&\overset{(a)}{\leq} (1+\tau_6)\mathbb{E}\left\|\mathbf{X}^{(mH)} - \overline{\mathbf{X}}^{(mH)}\right\|_F^2 + (1+\tau_6^{-1})(1+\tau_7)(1-\omega)(1+\tau_8)\mathbb{E}\left\|\mathbf{X}^{(mH)} - \widehat{\mathbf{X}}^{(mH)}\right\|_F^2 \\
&\qquad + \phi\eta^2\mathbb{E}\left\|\sum_{t'=mH}^{(m+1)H-1}\left(\beta\mathbf{V}^{(t')} + \nabla\boldsymbol{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})\right)\right\|_F^2,
\end{aligned} \tag{70}$$

where $\phi_3 = (1+\tau_6^{-1})\left((1+\tau_7^{-1}) + (1+\tau_7)(1-\omega)(1+\tau_8^{-1})\right)$, (a) follows from (65) for bounding $\mathbb{E}\left\|\widehat{\mathbf{X}}^{((m+1)H)} - \mathbf{X}^{(mH)}\right\|_F^2$. Observe that since we are bounding this quantity separately for (a), we can use different coefficients here. In the above bound on $\mathbb{E}\left\|\widehat{\mathbf{X}}^{((m+1)H)} - \mathbf{X}^{(mH)}\right\|_F^2$ from (65), instead of using the same τ_1, τ_2, we used τ_7, τ_8, respectively.

Substituting the above bound on T_3 into (69) and the substituting the resulting bound on T_2 from (69) and on T_1 from (68) into (67) gives

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{X}^{((m+1)H)} - \widehat{\mathbf{X}}^{((m+1)H)}\right\|_F^2 &\leq b_1\mathbb{E}\left\|\mathbf{X}^{(mH)} - \overline{\mathbf{X}}^{(mH)}\right\|_F^2 + b_2\mathbb{E}\left\|\mathbf{X}^{(mH)} - \widehat{\mathbf{X}}^{(mH)}\right\|_F^2 \\
&\quad + b_3\eta^2\mathbb{E}\left\|\sum_{t'=mH}^{(m+1)H-1}\left(\beta\mathbf{V}^{(t')} + \nabla\boldsymbol{F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})\right)\right\|_F^2,
\end{aligned} \tag{71}$$

where $b_1 = (1+\tau_3^{-1})\gamma^2\lambda^2(1+\tau_5)(1+\tau_6)$, $b_2 = (1+\tau_3)(1-\omega)(1+\tau_4) + (1+\tau_3^{-1})\gamma^2\lambda^2(1+\tau_5)(1+\tau_6^{-1})(1+\tau_7)(1-\omega)(1+\tau_8)$, $b_3 = (1+\tau_3)(1-\omega)(1+\tau_4^{-1}) + (1+\tau_3^{-1})\gamma^2\lambda^2(1+\tau_5)(1+\tau_6^{-1})\left((1+\tau_7^{-1}) + (1+\tau_7)(1-\omega)(1+\tau_8^{-1})\right) + (1+\tau_3^{-1})\gamma^2\lambda^2(1+\tau_5^{-1})$. □

C. Setting up parameters

We need to set the parameters such that we get $(1+\nu_1)\max\{a_1+b_1, a_2+b_2\} < 1$, this will give a contractive recursion in (36) and will lead to our convergence results. Recall the definitions of a_1, a_2 and b_1, b_2 from Lemma 7 and Lemma 8, respectively.

$$a_1 = (1+\alpha_5^{-1})(1+\alpha_1)(1-\gamma\delta)^2, \tag{72}$$
$$a_2 = (1+\alpha_5^{-1})(1+\alpha_1^{-1})\gamma^2\lambda^2(1+\tau_1)(1-\omega)(1+\tau_2), \tag{73}$$
$$b_1 = (1+\tau_3^{-1})\gamma^2\lambda^2(1+\tau_5)(1+\tau_6), \tag{74}$$
$$b_2 = (1+\tau_3)(1-\omega)(1+\tau_4) + (1+\tau_3^{-1})\gamma^2\lambda^2(1+\tau_5)(1+\tau_6^{-1})(1+\tau_7)(1-\omega)(1+\tau_8). \tag{75}$$

Here, ω, δ, λ are fixed parameters and are given to us. Among the rest, there is no trade-off when choosing $\alpha_5, \tau_1, \tau_2, \tau_4, \tau_5, \tau_7, \tau_8$, and we can chose them without any constraints. We need to carefully choose the remaining parameters $\alpha_1, \tau_3, \tau_6, \gamma$ as they contribute differently to different terms in the above equations. We will set all these parameters as follows:

$$\tau_i = \frac{\omega}{4}, \text{ for } i = 1,2,3,4,5,7,8; \quad \tau_6 = \frac{4}{\omega}; \tag{76}$$
$$\alpha_1 = \frac{\gamma\delta}{2}; \quad \alpha_5^{-1} = \frac{\gamma\delta}{2}; \quad \gamma^* = \frac{2\delta\omega^3}{(128\lambda^2 + 24\lambda^2\omega^2 + 4\delta^2\omega^2)}. \tag{77}$$

Now we substitute these values into (72)-(75).

- For a_1, we will use $\alpha_5^{-1} \leq \frac{\gamma\delta}{2}$ and $(1+\frac{\gamma\delta}{2})(1-\gamma\delta) \leq (1-\frac{\gamma\delta}{2})$ (since $\gamma\delta \leq 1$ which is true for $\gamma = \gamma^*$).

$$a_1 \leq (1+\frac{\gamma\delta}{2})^2(1-\gamma\delta)^2 \leq (1-\frac{\gamma\delta}{2})^2. \tag{78}$$

- For a_2, we will use $\alpha_5^{-1} \leq \frac{\omega}{4}$ (which holds because $\frac{\gamma\delta}{2} \leq \frac{\omega}{4}$ for $\gamma = \gamma^*$), $(1+\frac{\omega}{4})^3(1-\omega) \leq (1-\frac{\omega}{4})$, and $\frac{1}{\gamma\delta} \geq 1$.

$$a_2 \leq (1+\frac{\omega}{4})(1+\frac{2}{\gamma\delta})\gamma^2\lambda^2(1+\frac{\omega}{4})(1-\omega)(1+\frac{\omega}{4}) \leq \frac{3\gamma\lambda^2}{\delta}(1-\frac{\omega}{4}). \tag{79}$$

- For b_1, we will use $(1+\frac{4}{\omega}) \leq \frac{5}{\omega}$, $(1+\frac{\omega}{4}) \leq \frac{5}{4}$, and $\frac{125}{4} \leq 32$.

$$b_1 = (1+\frac{4}{\omega})\gamma^2\lambda^2(1+\frac{\omega}{4})(1+\frac{4}{\omega}) \leq \gamma^2\lambda^2\frac{25}{\omega^2}\frac{5}{4} \leq \gamma^2\lambda^2\frac{32}{\omega^2}. \tag{80}$$

- For b_2, we will use $(1+\frac{\omega}{4})^2(1-\omega) \leq (1+\frac{\omega}{4})^3(1-\omega) \leq (1-\frac{\omega}{4})$ in the first inequality, and $(1+\frac{4}{\omega}) \leq \frac{5}{\omega}$ and $(1+\frac{\omega}{4}) \leq \frac{5}{4}$ in the second inequality.

$$\begin{aligned} b_2 &= (1+\frac{\omega}{4})^2(1-\omega) + (1+\frac{4}{\omega})\gamma^2\lambda^2(1+\frac{\omega}{4})^4(1-\omega) \\ &\leq (1-\frac{\omega}{4}) + (1+\frac{4}{\omega})\gamma^2\lambda^2(1+\frac{\omega}{4})(1-\frac{\omega}{4}) \\ &\leq (1-\frac{\omega}{4})\left(1+\frac{5}{\omega}\gamma^2\lambda^2\frac{5}{4}\right) \\ &= (1-\frac{\omega}{4})\left(1+\gamma^2\lambda^2\frac{25}{4\omega}\right). \end{aligned} \tag{81}$$

Bounding (a_1+b_1). Adding the bounds in (78) and (80), we get

$$a_1 + b_1 \leq \underbrace{(1-\frac{\gamma\delta}{2})^2 + \gamma^2\lambda^2\frac{32}{\omega^2}}_{=:\ h_1(\gamma)}. \tag{82}$$

It can be verified that $h_1(\gamma)$ is a convex function in γ and attains minima at $\gamma' = \frac{2\delta\omega^2}{128\lambda^2+\delta^2\omega^2}$ with value $h_1(\gamma') = \frac{128\lambda^2}{128\lambda^2+\delta^2\omega^2} < 1$.

Putting this γ' in the expression for a_2+b_2 will not give a quantity that is less than one. In the following, we will derive a value of γ^* that works for both a_1+b_1 and a_2+b_2. Let $\gamma^* = s\gamma'$ for some $s \in [0,1]$. We will derive the value of s (and of γ^*).

By the convexity of h, we have

$$\begin{aligned} h_1(\gamma^*) = h_1(s\gamma') &= h_1((1-s)0 + s\gamma') \\ &\leq (1-s)h_1(0) + sh_1(\gamma') \\ &\leq (1-s) + s\frac{128\lambda^2}{128\lambda^2+\delta^2\omega^2} \\ &= 1 - s\frac{\delta^2\omega^2}{128\lambda^2+\delta^2\omega^2}. \end{aligned} \tag{83}$$

Bounding (a_2+b_2). Adding the bounds in (79) and (81) gives:

$$\begin{aligned} a_2 + b_2 &\leq (1-\frac{\omega}{4})\left(1+\frac{3\gamma\lambda^2}{\delta}+\gamma^2\lambda^2\frac{25}{4\omega}\right) \\ &\leq \underbrace{(1-\frac{\omega}{4}) + \left(\frac{3\gamma\lambda^2}{\delta}+\gamma^2\lambda^2\frac{25}{4\omega}\right)}_{=:\ h_2(\gamma)}. \end{aligned} \tag{84}$$

Putting $\gamma = \gamma^* = s\gamma' = \frac{2\delta\omega^2 s}{D}$, where $D = (128\lambda^2+\delta^2\omega^2)$, we get

$$\begin{aligned} h_2(\gamma^*) &\leq (1-\frac{\omega}{4}) + \left(3\lambda^2\frac{2\omega^2 s}{D} + \frac{25\lambda^2}{4\omega}\frac{4\delta^2\omega^4 s^2}{D^2}\right) \\ &\leq (1-\frac{\omega}{4}) + \frac{s}{D}\left(6\lambda^2\omega^2 + \frac{25\lambda^2\delta^2\omega^3 s}{D}\right) \\ &\leq (1-\frac{\omega}{4}) + \frac{s}{D}\left(6\lambda^2\omega^2 + 25\lambda^2\right) && \text{(Since } D \geq \delta^2\omega^2 \geq \delta^2\omega^3 s \text{ because } \omega, s \leq 1) \\ &\leq (1-\frac{\omega}{4}) + \frac{s}{D}\left(6\lambda^2\omega^2 + 32\lambda^2\right). \end{aligned} \tag{85}$$

Equating the upper bounds on $h_1(\gamma^*)$ and $h_2(\gamma^*)$, we get

$$\begin{aligned} 1 - s\frac{\delta^2\omega^2}{D} &= (1-\frac{\omega}{4}) + \frac{s}{D}\left(6\lambda^2\omega^2 + 32\lambda^2\right) \\ \iff \frac{\omega}{4} &= \frac{s}{D}(32\lambda^2 + 6\lambda^2\omega^2 + \delta^2\omega^2) \end{aligned}$$

$$\iff s = \frac{\omega D}{(128\lambda^2 + 24\lambda^2\omega^2 + 4\delta^2\omega^2)} < 1. \tag{86}$$

With this, we have $\gamma^* = s\gamma' = \frac{2\delta\omega^2 s}{D} = \frac{2\delta\omega^3}{(128\lambda^2+24\lambda^2\omega^2+4\delta^2\omega^2)}$.

Substituting the value of s from (86) into (83), we get

$$h_1(\gamma^*) \le 1 - \frac{\delta^2\omega^3}{(128\lambda^2 + 24\lambda^2\omega^2 + 4\delta^2\omega^2)} = 1 - \frac{\gamma^*\delta}{2}. \tag{87}$$

Thus we have

$$\max\{a_1 + b_1, a_2 + b_2\} \le \max\{h_1(\gamma^*), h_2(\gamma^*)\} \le 1 - \frac{\gamma^*\delta}{2}.$$

Taking $\nu_1 = \frac{\gamma^*\delta}{4}$ and using the inequality $(1 + x/2)(1 - x) \le (1 - x/2)$ (for $x = \frac{\gamma^*\delta}{2} \le 1$), we get

$$(1 + \nu_1)\max\{a_1 + b_1, a_2 + b_2\} \le 1 - \frac{\gamma^*\delta}{4} \le 1 - \frac{\delta^2\omega^3}{1224}, \tag{88}$$

where the last inequality follows by substituting the trivial upper bounds of $\lambda \le 2$ and $\delta, \omega \le 1$ in the denominator of the expression of γ^*.

Bounding $c_2 + c_4$ in (36).

$$c_2 = 2(1 + \nu_1)(a_{31} + a_{32}) + 2(1 + \nu_1^{-1}), \tag{89}$$
$$c_4 = 2(1 + \nu_1)(b_{31} + b_{32} + b_{33}) + 2(1 + \nu_1^{-1}), \tag{90}$$

where

$$a_{31} = (1 + \alpha_1)(1 - \gamma\delta)^2(1 + \alpha_5) + (1 + \alpha_1^{-1})\gamma^2\lambda^2(1 + \alpha_5), \tag{91}$$
$$a_{32} = (1 + \alpha_5^{-1})(1 + \alpha_1^{-1})\gamma^2\lambda^2\left((1 + \tau_1^{-1}) + (1 + \tau_1)(1 - \omega)(1 + \tau_2^{-1})\right), \tag{92}$$
$$b_{31} = (1 + \tau_3)(1 - \omega)(1 + \tau_4^{-1}), \tag{93}$$
$$b_{32} = (1 + \tau_3^{-1})\gamma^2\lambda^2(1 + \tau_5)(1 + \tau_6^{-1})\left((1 + \tau_7^{-1}) + (1 + \tau_7)(1 - \omega)(1 + \tau_8^{-1})\right), \tag{94}$$
$$b_{33} = (1 + \tau_3^{-1})\gamma^2\lambda^2(1 + \tau_5^{-1}). \tag{95}$$

Now we substituting the parameter setting from (76), (77) into the above equations.

- For a_{31}, we will use $(1 + \frac{\gamma\delta}{2})(1 - \gamma\delta)^2 \le (1 - \frac{\gamma\delta}{2})(1 - \gamma\delta) \le 1$ and $(1 + \frac{2}{\gamma\delta}) \le \frac{3}{\gamma\delta}$ (both follow from $\gamma\delta \le 1$).

$$\begin{aligned} a_{31} &= (1 + \frac{\gamma\delta}{2})(1 - \gamma\delta)^2(1 + \frac{2}{\gamma\delta}) + (1 + \frac{2}{\gamma\delta})^2\gamma^2\lambda^2 \\ &\le \frac{3}{\gamma\delta} + (\frac{3}{\gamma\delta})^2\gamma^2\lambda^2 = \frac{3}{\gamma\delta}\left(1 + \frac{3\gamma\lambda^2}{\delta}\right) \end{aligned} \tag{96}$$

- For a_{32}, we will use $(1 + \frac{\gamma\delta}{2}) \le \frac{3}{2}$, $(1 + \frac{2}{\gamma\delta}) \le \frac{3}{\gamma\delta}$, and $(1 + \frac{\omega}{4})(1 - \omega) \le (1 - \frac{3\omega}{4}) \le 1$ and $(1 + \frac{4}{\omega}) \le \frac{5}{\omega}$.

$$\begin{aligned} a_{32} &= (1 + \frac{\gamma\delta}{2})(1 + \frac{2}{\gamma\delta})\gamma^2\lambda^2\left((1 + \frac{4}{\omega}) + (1 + \frac{\omega}{4})(1 - \omega)(1 + \frac{4}{\omega})\right) \\ &\le \frac{3}{2}\frac{3}{\gamma\delta}\gamma^2\lambda^2\frac{10}{\omega} = \frac{45\gamma\lambda^2}{\delta\omega}. \end{aligned} \tag{97}$$

- For b_{31}, we will use $(1 + \frac{\omega}{4})(1 - \omega) \le (1 - \frac{3\omega}{4})$.

$$b_{31} = (1 + \frac{\omega}{4})(1 - \omega)(1 + \frac{4}{\omega}) \le (1 - \frac{3\omega}{4})(1 + \frac{4}{\omega}) \le \frac{4}{\omega} - 2. \tag{98}$$

- For b_{32}, we will use $(1 + \frac{4}{\omega}) \le \frac{5}{\omega}$, $(1 + \frac{\omega}{4}) \le \frac{5}{4}$, and $\left((1 + \frac{4}{\omega}) + (1 + \frac{\omega}{4})(1 - \omega)(1 + \frac{4}{\omega})\right) \le \frac{10}{\omega}$ as in a_{32}.

$$\begin{aligned} b_{32} &= (1 + \frac{4}{\omega})\gamma^2\lambda^2(1 + \frac{\omega}{4})(1 + \frac{\omega}{4})\left((1 + \frac{4}{\omega}) + (1 + \frac{\omega}{4})(1 - \omega)(1 + \frac{4}{\omega})\right) \\ &\le \frac{5}{\omega}\gamma^2\lambda^2(\frac{5}{4})^2\frac{10}{\omega} = \frac{625}{8}\frac{\gamma^2\lambda^2}{\omega^2} \le \frac{79\gamma^2\lambda^2}{\omega^2}. \end{aligned} \tag{99}$$

- For b_{33}, we will use

$$b_{33} = (1 + \frac{4}{\omega})\gamma^2\lambda^2(1 + \frac{4}{\omega}) \le \frac{25\gamma^2\lambda^2}{\omega^2}. \tag{100}$$

Substituting the bounds on a_{31}, a_{32} from (96), (97), respectively, and $\nu_1 = \frac{\gamma\delta}{4}$ (where $\gamma = \gamma^*$ is defined in (77)) into (89), we get:

$$c_2 \leq 2(1+\frac{\gamma\delta}{4})\left(\frac{3}{\gamma\delta}\left(1+\frac{3\gamma\lambda^2}{\delta}\right)+\frac{45\gamma\lambda^2}{\delta\omega}\right)+2(1+\frac{4}{\gamma\delta}). \tag{101}$$

Similarly, substituting the bounds on b_{31}, b_{32}, b_{33} from (98), (99), (100), respectively, and $\nu_1 = \frac{\gamma\delta}{4}$ (where $\gamma = \gamma^*$ is defined in (77)) into (90), we get:

$$c_4 \leq 2(1+\frac{\gamma\delta}{4})\left(\frac{4}{\omega}-2+\frac{104\gamma^2\lambda^2}{\omega^2}\right)+2(1+\frac{4}{\gamma\delta}). \tag{102}$$

Adding the bounds on c_2 and c_4 gives

$$c_2+c_4 \leq 2(1+\frac{\gamma\delta}{4})\left(\frac{3}{\gamma\delta}+\frac{9\lambda^2}{\delta^2}+\frac{45\gamma\lambda^2}{\delta\omega}+\frac{104\gamma^2\lambda^2}{\omega^2}+\frac{4}{\omega}-2\right)+4(1+\frac{4}{\gamma\delta}). \tag{103}$$

Putting the bounds from (88) and (103) back into (36), we get

$$\begin{aligned}S^{(t)} \leq \left(1-\frac{\gamma\delta}{4}\right)S^{(mH)} &+ 2c_1\eta^2H^2n\left(2(M^2+1)G^2+\sigma^2\right)+c_1\eta^2H\beta^2\sum_{t'=mH}^{t-1}\mathbb{E}\left\|\mathbf{V}^{(t')}\right\|_F^2 \\ &+ 2c_1\eta^2H(M^2+1)L^2\sum_{t'=mH}^{t-1}S^{(t')}+2c_1\eta^2H(M^2+1)nB^2\sum_{t'=mH}^{t-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|_2^2,\end{aligned} \tag{104}$$

where $c_1 = c_2 + c_4$ and the bound on $c_2 + c_4$ is given in (103), and $\gamma = \gamma^*$ is defined in (77).

D. Omitted Details from Section V-E

Proof of Proposition 3.

$$\begin{aligned}
&\mathbb{E}\left\|\nabla \boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})\right\|_F^2 = \mathbb{E}\left\|\nabla f(\mathbf{X}^{(t')})\right\|_F^2+\mathbb{E}\left\|\nabla \boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')})-\nabla f(\mathbf{X}^{(t')})\right\|_F^2 \\
&=\mathbb{E}\left\|\nabla f(\mathbf{X}^{(t')})\right\|_F^2+\mathbb{E}\sum_{i=1}^n\left\|\nabla F(\mathbf{x}_i^{(t')},\xi_i^{(t')})-\nabla f(\mathbf{x}_i^{(t')})\right\|_2^2 \\
&\overset{(a)}{\leq}\mathbb{E}\left\|\nabla f(\mathbf{X}^{(t')})\right\|_F^2+n\sigma^2+M^2\mathbb{E}\left\|\nabla f(\mathbf{X}^{(t')})\right\|_F^2 \\
&=(M^2+1)\mathbb{E}\left\|\nabla f(\mathbf{X}^{(t')})\right\|_F^2+n\sigma^2 \\
&=(M^2+1)\mathbb{E}\left\|\nabla f(\mathbf{X}^{(t')})-\nabla f(\overline{\mathbf{X}}^{(t')})+\nabla f(\overline{\mathbf{X}}^{(t')})\right\|_F^2+n\sigma^2 \qquad (\text{Where } \nabla f(\overline{\mathbf{X}}^{(t')}) = [\nabla f_1(\overline{\mathbf{x}}^{(t')})\dots\nabla f_n(\overline{\mathbf{x}}^{(t')})]) \\
&\leq 2(M^2+1)\left(\mathbb{E}\left\|\nabla f(\mathbf{X}^{(t')})-\nabla f(\overline{\mathbf{X}}^{(t')})\right\|_F^2+\mathbb{E}\left\|\nabla f(\overline{\mathbf{X}}^{(t')})\right\|_F^2\right)+n\sigma^2 \\
&\overset{(b)}{\leq} 2(M^2+1)\left(L^2\mathbb{E}\left\|\mathbf{X}^{(t')}-\overline{\mathbf{X}}^{(t')}\right\|_F^2+\mathbb{E}\sum_{i=1}^n\left\|\nabla f_i(\overline{\mathbf{x}}^{(t')})\right\|_2^2\right)+n\sigma^2 \\
&\overset{(c)}{\leq} 2(M^2+1)\left(L^2\mathbb{E}\left\|\mathbf{X}^{(t')}-\overline{\mathbf{X}}^{(t')}\right\|_F^2+nG^2+nB^2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|_2^2\right)+n\sigma^2 \\
&=2(M^2+1)\left(L^2\Xi^{(t')}+nG^2+nB^2\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t')})\right\|_2^2\right)+n\sigma^2
\end{aligned}$$

where (a) follows from Assumption 2, (b) follows from the L-smoothness of f, and (c) follows from Assumption 3. □

E. Omitted Details from Section V-G

Claim 1. *We have* $\left(1-\frac{\alpha}{4}\right)^{\lfloor\frac{t-j}{H}\rfloor} \leq 2\left(1-\frac{\alpha}{8H}\right)^{t-j}$.

Proof. First note that $(1-\frac{\alpha}{4})^{1/H} \leq \exp(-\frac{\alpha}{4H}) \leq 1-\frac{\alpha}{8H}$ and also that $\lfloor\frac{t-j}{H}\rfloor \geq \frac{t-j}{H}-1$.

$$\begin{aligned}\left(1-\frac{\alpha}{4}\right)^{\lfloor\frac{t-j}{H}\rfloor} &= \left[\left(1-\frac{\alpha}{4}\right)^{1/H}\right]^{H\lfloor\frac{t-j}{H}\rfloor} \leq \left(1-\frac{\alpha}{8H}\right)^{H\lfloor\frac{t-j}{H}\rfloor} \\ &\leq \left(1-\frac{\alpha}{8H}\right)^{t-j}\left(1-\frac{\alpha}{8H}\right)^{-H} \leq 2\left(1-\frac{\alpha}{8H}\right)^{t-j}.\end{aligned}$$

In the last inequality we used $\left(1-\frac{\alpha}{8H}\right)^{-H} \leq 2$, which can be shown as follows:

$$\left(1-\frac{\alpha}{8H}\right)^{-H} = \left(\frac{1}{1-\frac{\alpha}{8H}}\right)^{H} \overset{\text{(a)}}{\leq} \left(1+\frac{\alpha}{4H}\right)^{H} \leq \exp(\frac{\alpha}{4}) \leq 2,$$

where (a) holds because $\frac{\alpha}{8H} \leq \frac{1}{2}$. □

F. Completing the Convergence Proof

Note that $\Xi^{(t)} \sum_{i=1}^{n} \mathbb{E}\left\|\mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)}\right\|_2^2 \leq S^{(t)}$ for any $t \in [T]$. Substituting this and the bound from (50) in the last term of (26), we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|_2^2 \leq \frac{16(1-\beta)(f(\overline{\mathbf{x}}^{(0)})-f^*)}{\eta T} + \frac{16\eta L}{(1-\beta)}\Big(\frac{\sigma^2+2(M^2+n)G^2}{n}\Big) + \eta^2\frac{128L^2J_1}{n} + \eta^2\frac{128L^2J_2}{n}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|^2. \tag{105}$$

where $J_1 = \left(\frac{8A\eta^2}{\alpha} + \left(\frac{32DH}{\alpha}\right)\left(\frac{2(M^2+1)nG^2+n\sigma^2}{(1-\beta)}\right)\right)$ and $J_2 = \left(\frac{32CH}{\alpha} + \left(\frac{32DH}{\alpha}\right)\frac{2(M^2+1)nB^2}{(1-\beta)}\right)$, $A = 2c_1H^2n\left(2(M^2+1)G^2+\sigma^2\right)$, $C = 2c_1H(M^2+1)nB^2$, and $D = \frac{c_1H\beta^2}{(1-\beta)}$ and c_1 defined below. If $\eta \leq \sqrt{\frac{n}{256L^2J_2}}$, then taking the last term on the LHS gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|_2^2 \leq \frac{32(1-\beta)(f(\overline{\mathbf{x}}^{(0)})-f^*)}{\eta T} + \frac{32\eta L}{(1-\beta)}\Big(\frac{\sigma^2+2(M^2+n)G^2}{n}\Big) + \eta^2\frac{256L^2J_1}{n}. \tag{106}$$

Choosing $\eta = (1-\beta)\sqrt{\frac{n}{T}}$ and running the algorithm for $T \geq \max\{U_1, U_2, U_3, U_4, U_5\}$ iterations completes the proof of Theorem 1.
Here, $U_1 = \frac{81n\beta^8}{4(1-\beta)^4}$, $U_2 = \frac{9(M^2+n)\beta^4L^2}{4(1-\beta^2)}$, $U_3 = \frac{72(M^2+n)\beta^2L^2B^2}{(1-\beta)^2}$, $U_4 = 256L^2J_2(1-\beta)^2$ and $U_5 = \frac{512DH(M^2+1)L^2(1-\beta)n}{\delta\gamma}$, with $J_2 = \frac{128CH}{\gamma\delta} + \left(\frac{128DH}{\gamma\delta}\right)\left(\frac{2(M^2+1)nB^2}{1-\beta}\right)$, $D = \frac{c_1H\beta^2}{(1-\beta)}$,
$C = 2c_1H(M^2+1)nB^2$ and $c_1 = 2(1+\frac{\gamma\delta}{4})\left(\frac{3}{\gamma\delta} + \frac{9\lambda^2}{\delta^2} + \frac{45\gamma\lambda^2}{\delta\omega} + \frac{104\gamma^2\lambda^2}{\omega^2} + \frac{4}{\omega} - 2\right)$.

APPENDIX D
PRELIMINARIES FOR CONVERGENCE WITH RELAXED ASSUMPTIONS

Fact 5. *Consider the variance bound on the stochastic gradient for nodes $i \in [n]$:*

$$\mathbb{E}_{\xi_i}\left\|\nabla F_i(\mathbf{x},\xi_i) - \nabla f_i(\mathbf{x})\right\|^2 \leq \sigma_i^2,$$

where $\mathbb{E}_{\xi_i}[\nabla F_i(\mathbf{x},\xi_i)] = \nabla f_i(\mathbf{x})$, then:

$$\mathbb{E}_{\boldsymbol{\xi}^{(t)}}\left\|\frac{1}{n}\sum_{j=1}^{n}\left(\nabla f_j(\mathbf{x}_j^{(t)}) - \nabla F_j(\mathbf{x}_j^{(t)},\xi_j^{(t)})\right)\right\|^2 \leq \frac{\bar{\sigma}^2}{n} \tag{107}$$

where $\boldsymbol{\xi}^{(t)} = \{\xi_1^{(t)}, \xi_2^{(t)}, \ldots, \xi_n^{(t)}\}$ denotes the stochastic sample for the nodes at any timestep t and $\frac{\sum_{j=1}^{n}\sigma_j^2}{n} = \bar{\sigma}^2$

Proof.

$$\mathbb{E}_{\xi^{(t)}}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla f_j(\mathbf{x}_j^{(t)}) - \frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\mathbf{x}_j^{(t)},\xi_j^{(t)})\right\|^2 = \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}_{\xi^{(t)}}\|\nabla f_j(\mathbf{x}_j^{(t)}) - \nabla F_j(\mathbf{x}_j^{(t)},\xi_j^{(t)})\|^2 + \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}_{\xi^{(t)}}\left\langle\nabla f_i(\mathbf{x}_i^{(t)}) - \nabla F_i(\mathbf{x}_i^{(t)},\xi_j^{(t)}), \nabla f_j(\mathbf{x}_j^{(t)}) - \nabla F_j(\mathbf{x}_j^{(t)},\xi_j^{(t)})\right\rangle$$

Since ξ_i is independent of ξ_j, the second term is zero in expectation, thus the above reduces to:

$$\mathbb{E}_{\xi^{(t)}}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla f_j(\mathbf{x}_j^{(t)}) - \frac{1}{n}\sum_{j=1}^{n}\nabla F_j(\mathbf{x}_j^{(t)},\xi_j^{(t)})\right\|^2 = \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}_{\xi^{(t)}}\|\nabla f_j(\mathbf{x}_j^{(t)}) - \nabla F_j(\mathbf{x}_j^{(t)},\xi_j^{(t)})\|^2$$

$$\leq \frac{1}{n^2}\sum_{j=1}^{n}\sigma_j^2 = \frac{\bar{\sigma}^2}{n}$$

□

Fact 6. *Consider the set of synchronization indices $\{I_{(1)}, I_{(2)}, \ldots, I_{(k)}, \ldots\} \in \mathcal{I}_T$. We assume that the maximum gap between any two consecitive elements in $\mathcal{I}_T$ is bounded by H. Let $\xi^{(t)} = \{\xi_1^{(t)}, \xi_2^{(t)}, \ldots, \xi_n^{(t)}\}$ denote the stochastic samples for the nodes at any timestep t. Consider any two consecutive synchronization indices $I_{(k)}$ and $I_{(k+1)}$, then for learning rate η, we have:*

$$\mathbb{E}\left[\left\|\sum_{t'=I_{(k)}}^{I_{(k+1)}-1}\eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}))\right\|_F^2\right] \leq 2nH^2G^2\eta^2\left(1 + \frac{\beta^2}{(1-\beta)^2}\right). \tag{108}$$

Proof. Using the fact that the sequence gap is bounded by H, we have $I_{(t+1)} - I_{(t)} \leq H$ for all synchronization indices $I_{(t)} \in \mathcal{I}_T$. Thus we have:

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{t'=I_{(k)}}^{I_{(k+1)}-1}\eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}))\right\|_F^2\right] &\leq H\eta^2\sum_{t'=I_{(k)}}^{I_{(k+1)}-1}\mathbb{E}\left\|\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})\right\|_F^2 \\ &\leq 2H\eta^2\sum_{t'=I_{(k)}}^{I_{(k+1)}-1}\left[\mathbb{E}\left\|\beta\mathbf{V}^{(t')}\right\|_F^2 + \mathbb{E}\left\|\boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})\right\|_F^2\right]\end{aligned}$$

Using the bounded gradient assumption and definition of gap H, we can bound the above as:

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{t'=I_{(k)}}^{I_{(k+1)}-1}\eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}))\right\|_F^2\right] &\leq 2H\eta^2\beta^2\sum_{t'=I_{(k)}}^{I_{(k+1)}-1}\mathbb{E}\left\|\mathbf{V}^{(t')}\right\|_F^2 + 2nH^2G^2\eta^2 \\ &= 2H\eta^2\beta^2\sum_{t'=I_{(k)}}^{I_{(k+1)}-1}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{v}_i^{(t')}\right\|^2 + 2nH^2G^2\eta^2\end{aligned} \tag{109}$$

Now we show that $\mathbb{E}\left\|\mathbf{v}_i^{(t)}\right\|^2 \leq \frac{G^2}{(1-\beta)^2}$ for all $i \in [n]$ and for every $t \geq 0$. Fix an arbitrary $i \in [n]$ and $t \geq 0$. Define $\theta_t = \sum_{k=0}^{t}\beta^k$, we then have:

$$\begin{aligned}\mathbb{E}\left\|\mathbf{v}_i^{(t)}\right\|^2 &= \theta_t^2\mathbb{E}\left\|\sum_{k=0}^{t}\frac{\beta^{t-k}}{\theta_t}\nabla F(\mathbf{x}_i^{(k)}, \xi_i^{(k)})\right\|^2 \\ &\leq \theta_t\sum_{k=0}^{t}\beta^{t-k}\mathbb{E}\left\|\nabla F(\mathbf{x}_i^{(k)}, \xi_i^{(k)})\right\|^2 \\ &\leq \theta_t\sum_{k=0}^{t}\left[\beta^{t-k}G^2\right] \\ &= G^2\theta_t^2\end{aligned}$$

Here the first inequality follows from the Jensen's inequality and the second inequality follows from the bounded gradient assumption. We now note the following bound for θ_t:

$$\theta_t = \sum_{k=0}^{t}\beta^k \leq \sum_{k=0}^{\infty}\beta^k \leq \frac{1}{(1-\beta)}$$

Thus, for all t and all $i \in [n]$, we have:

$$\mathbb{E}\left\|\mathbf{v}_i^{(t)}\right\|^2 \leq \frac{G^2}{(1-\beta)^2} \tag{110}$$

Substituting the bound $\mathbb{E}\|\mathbf{v}_i^{(t)}\|^2 \leq \frac{G^2}{(1-\beta)^2}$ in (109) gives

$$\mathbb{E}\left[\left\|\sum_{t'=I_{(k)}}^{I_{(k+1)}-1}\eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}))\right\|_F^2\right] \leq 2H^2\eta^2\beta^2n\frac{G^2}{(1-\beta)^2} + 2nH^2G^2\eta^2.$$

This completes the proof of Fact 6. □

Fact 7 (Triggering rule, [19]). *Consider the set of nodes $\Gamma^{(t)}$ which do not communicate at time t. For a threshold sequence $\{c_t\}_{t=0}^{T-1}$, the triggering rule in Algorithm 1 dictates that*

$$\|\mathbf{x}_i^{(t+\frac{1}{2})} - \hat{\mathbf{x}}_i^{(t)}\|^2 \leq c_t \eta^2 \qquad \forall i \in \Gamma^{(t)}.$$

Using the matrix notation, this implies that:

$$\left\|(\mathbf{X}^{(t+\frac{1}{2})} - \hat{\mathbf{X}}^{(t)})(\mathbf{I} - \mathbf{P}^{(t)})\right\|_F^2 \leq n c_t \eta^2. \tag{111}$$

Fact 8 (Lemma 16, [21]). *For doubly stochastic matrix $\mathbf{W}$ with second largest eigenvalue $1-\delta = |\lambda_2(\mathbf{W})| < 1$, we have:*

$$\left\|\mathbf{W}^k - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right\| = (1-\delta)^k \tag{112}$$

for any non-negative integer k.

Claim 2. *For any $n \in \mathbb{N}$, we have $\left\|\frac{\mathbf{1}\mathbf{1}^T}{n} - \mathbf{I}\right\|_2 = 1$ where $\mathbf{1} = [1\,1\,\ldots\,1]_{1\times n}^T$*

Proof. Note that $\frac{\mathbf{1}\mathbf{1}^T}{n}$ is a symmetric doubly stochastic matrix with eigenvalues 1 and 0 (with algebraic multiplicity $n-1$). Thus, it has the eigen-decomposition $\frac{\mathbf{1}\mathbf{1}^T}{n} = \mathbf{U}\mathbf{D}\mathbf{U}^T$ where columns of $\mathbf{U}$ are orthogonal and $\mathbf{D} = diag([1\,0\ldots0])$, which gives us:

$$\left\|\frac{\mathbf{1}\mathbf{1}^T}{n} - \mathbf{I}\right\|_2 = \left\|\mathbf{U}\mathbf{D}\mathbf{U}^T - \mathbf{U}\mathbf{U}^T\right\|_2 = \|\mathbf{D}-\mathbf{I}\|_2 = \left\|\begin{bmatrix}1&0&\ldots&0\\0&0&\ldots&0\\\vdots&\vdots&\ddots&0\\0&0&\ldots&0\end{bmatrix} - \begin{bmatrix}1&0&\ldots&0\\0&1&\ldots&0\\\vdots&\vdots&\ddots&0\\0&0&\ldots&1\end{bmatrix}\right\|_2 = 1$$

□

Appendix E
Proof of Theorem 2 (Non-convex objective)

From the recurrence relation of the virtual sequence (15), we have:

$$\begin{aligned}
\mathbb{E}_{\xi_{(t)}}[f(\widetilde{\mathbf{x}}^{(t+1)})] &= \mathbb{E}_{\xi_{(t)}} f\left(\widetilde{\mathbf{x}}^{(t)} - \frac{\eta}{(1-\beta)}\frac{1}{n}\sum_{i=1}^n \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right)\\
&\leq f(\widetilde{\mathbf{x}}^{(t)}) - \left\langle \nabla f(\widetilde{\mathbf{x}}^{(t)}), \frac{\eta}{(1-\beta)}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\xi_{(t)}}[\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})]\right\rangle\\
&\quad + \frac{L}{2}\frac{\eta^2}{(1-\beta)^2}\mathbb{E}_{\xi_{(t)}}\left\|\frac{1}{n}\sum_{i=1}^n \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\|^2\\
&\leq f(\widetilde{\mathbf{x}}^{(t)}) - \left\langle \nabla f(\widetilde{\mathbf{x}}^{(t)}), \frac{\eta}{(1-\beta)}\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)})\right\rangle + \frac{L}{2}\frac{\eta^2}{(1-\beta)^2}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)})\right\|^2\\
&\quad + \frac{L}{2}\frac{\eta^2}{(1-\beta)^2}\mathbb{E}_{\xi_{(t)}}\left\|\frac{1}{n}\sum_{i=1}^n (\nabla f_i(\mathbf{x}_i^{(t)}) - \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\|^2\\
&\leq f(\widetilde{\mathbf{x}}^{(t)}) - \left\langle \nabla f(\widetilde{\mathbf{x}}^{(t)}), \frac{\eta}{(1-\beta)}\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)})\right\rangle + \frac{L}{2}\frac{\eta^2}{(1-\beta)^2}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)})\right\|^2\\
&\quad + \frac{L\eta^2\bar{\sigma}^2}{2n(1-\beta)^2}
\end{aligned} \tag{113}$$

We now focus on bounding the second term in (113). First, note the following:

$$\begin{aligned}
\left\langle \nabla f(\widetilde{\mathbf{x}}^{(t)}), \frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)})\right\rangle &= \left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 - \left\langle \frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)}) - \nabla f(\widetilde{\mathbf{x}}^{(t)}), \frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)})\right\rangle\\
&= \left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 - \left\langle \frac{1}{n}\sum_{i=1}^n (\nabla f_i(\mathbf{x}_i^{(t)}) - \nabla f_i(\widetilde{\mathbf{x}}^{(t)})), \frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^{(t)})\right\rangle
\end{aligned}$$

$$\geq \frac{1}{2}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 - \frac{L^2}{2n}\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)} - \widetilde{\mathbf{x}}^{(t)}\right\|^2 \tag{114}$$

where in the last inequality, we've used the fact that $2\langle \mathbf{a}, \mathbf{b}\rangle \leq \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2$ for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ and the $L-$smoothness assumption for objectives $\{f_i\}_{i=1}^n$. We now state how to bound the last term on R.H.S. of (114). First, note the bound:

$$\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)} - \widetilde{\mathbf{x}}^{(t)}\right\|^2 \leq 2\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)}\right\|^2 + 2\sum_{i=1}^{n}\left\|\overline{\mathbf{x}}^{(t)} - \widetilde{\mathbf{x}}^{(t)}\right\|^2 \tag{115}$$

Using Lemma 5 to bound the second term in (115), we get:

$$\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)} - \widetilde{\mathbf{x}}^{(t)}\right\|^2 \leq 2\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)}\right\|^2 + \frac{2n\beta^4\eta^2}{(1-\beta)^3}\sum_{\tau=0}^{t-1}\left[\beta^{t-\tau-1}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(\tau)}, \xi_i^{(\tau)})\right\|^2\right] \tag{116}$$

Using the bound (116) in (114) and substituting it in (113), we have the following bound:

$$\begin{aligned}\mathbb{E}_{\xi_{(t)}}[f(\widetilde{\mathbf{x}}^{(t+1)})] \leq f(\widetilde{\mathbf{x}}^{(t)}) &+ \frac{L\eta^2\bar{\sigma}^2}{2n(1-\beta)^2} + \frac{L\eta^2}{2(1-\beta)^2}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 - \frac{\eta}{2(1-\beta)}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 \\ &+ \frac{\eta}{(1-\beta)}\frac{L^2}{n}\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)}\right\|^2 + \frac{L^2\eta^3\beta^4}{(1-\beta)^4}\sum_{\tau=0}^{t-1}\left[\beta^{t-\tau-1}\mathbb{E}_{\xi_{(t)}}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(\tau)}, \xi_i^{(\tau)})\right\|^2\right]\end{aligned}$$

Rearranging the terms, we can write:

$$\begin{aligned}\left(\frac{\eta}{2(1-\beta)} - \frac{L\eta^2}{2(1-\beta)^2}\right)\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 &\leq f(\widetilde{\mathbf{x}}^{(t)}) - \mathbb{E}_{\xi_{(t)}}f(\widetilde{\mathbf{x}}^{(t+1)}) + \frac{L\eta^2\bar{\sigma}^2}{2n(1-\beta)^2} \\ &+ \frac{L^2\eta}{(1-\beta)n}\sum_{i=1}^{n}\left\|\mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)}\right\|^2 + \frac{L^2\eta^3\beta^4}{(1-\beta)^4}\sum_{\tau=0}^{t-1}\left[\beta^{t-\tau-1}\mathbb{E}_{\xi_{(t)}}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(\tau)}, \xi_i^{(\tau)})\right\|^2\right]\end{aligned}$$

Summing from $t = 0$ to T gives us:

$$\begin{aligned}&\left(\frac{\eta}{2(1-\beta)} - \frac{L\eta^2}{2(1-\beta)^2}\right)\sum_{t=0}^{T-1}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 \\ &\leq f(\widetilde{\mathbf{x}}^{(0)}) - \mathbb{E}_{\xi_{(t)}}f(\widetilde{\mathbf{x}}^{(T)}) + \frac{L\eta^2\bar{\sigma}^2T}{2n(1-\beta)^2} + \frac{L^2\eta}{(1-\beta)n}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)}\right\|^2 \\ &\qquad + \frac{L^2\eta^3\beta^4}{(1-\beta)^4}\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\left[\beta^{t-\tau-1}\mathbb{E}_{\xi_{(t)}}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(\tau)}, \xi_i^{(\tau)})\right\|^2\right]\end{aligned}$$

Using the fact that $\mathbb{E}_{\xi_{(t)}}[\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})] = \nabla f_i(\mathbf{x}_i^{(t)})$ for all $i \in [n]$ and for all $t \in [T]$, we have $\mathbb{E}_{\xi_{(t)}}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\|^2 = \mathbb{E}_{\xi_{(t)}}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})\right\|^2 + \mathbb{E}_{\xi_{(t)}}\left\|\frac{1}{n}\sum_{i=1}^{n}(\nabla f_i(\mathbf{x}^{(t)}) - \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}))\right\|^2$. Using this equation along with the variance bound (107) from Fact 5, the fact that $\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\beta^{t-\tau-1} \leq {}^{T}/_{1-\beta}$ for $\beta \in (0, 1)$ and taking expectation w.r.t. the entire process:

$$\begin{aligned}&\leq f(\widetilde{\mathbf{x}}^{(0)}) - \mathbb{E}f(\widetilde{\mathbf{x}}^{(T)}) + \frac{L\eta^2\bar{\sigma}^2T}{2n(1-\beta)^2} + \frac{L^2\eta}{(1-\beta)n}\sum_{t=0}^{T-1}\sum_{i=1}^{n}\mathbb{E}\left\|\mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)}\right\|^2 \\ &\qquad + \frac{L^2\eta^3\beta^4\bar{\sigma}^2T}{n(1-\beta)^5} + \frac{L^2\eta^3\beta^4}{(1-\beta)^4}\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\left[\beta^{t-\tau-1}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(\tau)})\right\|^2\right]\end{aligned} \tag{117}$$

To bound the last term in (117), we note that:

$$\begin{aligned}\sum_{t=0}^{T-1}\sum_{\tau=0}^{t-1}\beta^{t-\tau-1}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(\tau)})\right\|^2 &= \sum_{\tau=0}^{T-2}\sum_{t=\tau+1}^{T-1}\beta^{t-\tau-1}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(\tau)})\right\|^2 \\ &\leq \frac{1}{(1-\beta)}\sum_{\tau=0}^{T-2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(\tau)})\right\|^2\end{aligned}$$

$$\leq \frac{1}{(1-\beta)} \sum_{t=0}^{T-1} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i^{(t)}) \right\|^2$$

Substituting the above bound in (117) and rearranging terms, we finally get:

$$\left(\frac{\eta}{2(1-\beta)} - \frac{L\eta^2}{2(1-\beta)^2} - \frac{L^2\eta^3\beta^4}{(1-\beta)^5} \right) \sum_{t=0}^{T-1} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i^{(t)}) \right\|^2$$
$$\leq f(\widetilde{\mathbf{x}}^{(0)}) - \mathbb{E} f(\widetilde{\mathbf{x}}^{(T)}) + \frac{L\eta^2\bar{\sigma}^2 T}{2n(1-\beta)^2} + \frac{L^2\eta}{(1-\beta)n} \sum_{t=0}^{T-1} \sum_{i=1}^{n} \mathbb{E} \left\| \mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)} \right\|^2 + \frac{L^2\eta^3\beta^4\bar{\sigma}^2 T}{n(1-\beta)^5} \tag{118}$$

If we select $\eta \leq \min\left\{ \frac{(1-\beta)}{4L}, \frac{(1-\beta)^2}{2\sqrt{2}L\beta^2} \right\}$, it can be shown that $\left(\frac{\eta}{2(1-\beta)} - \frac{L\eta^2}{2(1-\beta)^2} - \frac{L^2\eta^3\beta^4}{(1-\beta)^5} \right) \geq \frac{\eta}{4(1-\beta)}$. This gives:

$$\frac{\eta}{4(1-\beta)} \sum_{t=0}^{T-1} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i^{(t)}) \right\|^2 \leq f(\widetilde{\mathbf{x}}^{(0)}) - \mathbb{E}[f(\widetilde{\mathbf{x}}^{(T)})] + \frac{L\eta^2\bar{\sigma}^2 T}{2n(1-\beta)^2} + + \frac{L^2\eta^3\beta^4\bar{\sigma}^2 T}{n(1-\beta)^5}$$
$$+ \frac{L^2\eta}{(1-\beta)n} \sum_{t=0}^{T-1} \sum_{i=1}^{n} \mathbb{E} \left\| \mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)} \right\|^2$$

Multiplying both sides by $\frac{4(1-\beta)}{\eta T}$ and noting that $\mathbb{E}[f(\widetilde{\mathbf{x}}^{(T)})] \geq f^*$, we have:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i^{(t)}) \right\|^2 \leq \frac{4(1-\beta)}{\eta} \frac{(f(\mathbf{x}^{(0)}) - f^*)}{T} + \frac{2L\eta\bar{\sigma}^2}{n(1-\beta)}$$
$$+ \frac{4L^2}{nT} \sum_{t=0}^{T-1} \sum_{i=1}^{n} \mathbb{E} \left\| \mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)} \right\|^2 + \frac{4L^2\eta^2\beta^4\bar{\sigma}^2}{n(1-\beta)^4} \tag{119}$$

Now consider the time average of gradients evaluated at the global average $\overline{\mathbf{x}}^{(t)}$:

$$\begin{aligned}
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \nabla f(\overline{\mathbf{x}}^{(t)}) \right\|^2 &= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\overline{\mathbf{x}}^{(t)}) \right\|^2 \\
&= \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} (\nabla f_i(\overline{\mathbf{x}}^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})) + \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i^{(t)}) \right\|^2 \\
&\leq \frac{2}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} (\nabla f_i(\overline{\mathbf{x}}^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})) \right\|^2 + \frac{2}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i^{(t)}) \right\|^2 \\
&\leq \frac{2L^2}{nT} \sum_{t=0}^{T-1} \sum_{i=1}^{n} \mathbb{E} \left\| \overline{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)} \right\|^2 + \frac{2}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i^{(t)}) \right\|^2
\end{aligned} \tag{120}$$

where in the first inequality follows from Jensen's inequality and the second inequality follows from the $L-$smoothness assumption. We can bound the last term in (120) using (119) which gives us:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \nabla f(\overline{\mathbf{x}}^{(t)}) \right\|^2 \leq \frac{8(1-\beta)}{\eta} \frac{(f(\mathbf{x}^{(0)}) - f^*)}{T} + \frac{4L\eta\bar{\sigma}^2}{n(1-\beta)}$$
$$+ \left(\frac{8L^2}{nT} + \frac{2L^2}{nT} \right) \sum_{t=0}^{T-1} \sum_{i=1}^{n} \mathbb{E} \left\| \mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)} \right\|^2 + \frac{8L^2\eta^2\beta^4\bar{\sigma}^2}{n(1-\beta)^4} \tag{121}$$

Note that in our matrix form, $\mathbb{E} \left\| \bar{\mathbf{X}}^{(t)} - \mathbf{X}^{(t)} \right\|_F^2 = \sum_{i=1}^{n} \mathbb{E} \left\| \mathbf{x}_i^{(t)} - \overline{\mathbf{x}}^{(t)} \right\|^2$. Let $I_{(t+1)_0} \in \mathcal{I}_T$ denote the latest synchronization step before or equal to $(t+1)$. Then we have:

$$\mathbf{X}^{(t+1)} = \mathbf{X}^{I_{(t+1)_0}} - \textstyle\sum_{t'=I_{(t+1)_0}}^{t} \eta(\beta \mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}))$$
$$\bar{\mathbf{X}}^{(t+1)} = \bar{\mathbf{X}}^{I_{(t+1)_0}} - \textstyle\sum_{t'=I_{(t+1)_0}}^{t} \eta(\beta \mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})) \frac{\mathbf{1}\mathbf{1}^T}{n}$$

Thus the following holds:

$$\mathbb{E} \|\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t+1)}\|_F^2 = \mathbb{E} \left\| \mathbf{X}^{I_{(t+1)_0}} - \bar{\mathbf{X}}^{I_{(t+1)_0}} - \textstyle\sum_{t'=I_{(t+1)_0}}^{t} \eta(\beta \mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})) \left(\mathbf{I} - \frac{1}{n} \mathbf{1}\mathbf{1}^T \right) \right\|_F^2$$

$$\leq 2\mathbb{E}\|\mathbf{X}^{I_{(t+1)_0}}-\bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2+2\mathbb{E}\left\|\sum_{t'=I_{(t+1)_0}}^{t}\eta(\beta\mathbf{V}^{(t')}+\boldsymbol{\nabla}\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')}))\left(\mathbf{I}-\tfrac{1}{n}\mathbf{1}\mathbf{1}^T\right)\right\|_F^2$$

Using $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_2$ to split the second term in R.H.S. of above along with (112) from Fact 3 (with $k=0$) and further using the bound (108), we get:

$$\mathbb{E}\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t+1)}\|_F^2 \leq 2\mathbb{E}\|\mathbf{X}^{I_{(t+1)_0}}-\bar{\mathbf{X}}^{I_{(t+1)_0}}\|_F^2+4\eta^2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right) \tag{122}$$

We bound the first term in R.H.S. of (122) by Lemma 12 stated below and proved in Appendix G.

Lemma 12. ***(Consensus)*** *Let $\{\mathbf{x}_t^{(i)}\}_{t=0}^{T-1}$ be generated according to Algorithm 1 under assumptions of Theorem 2 with constant stepsize η, a threshold sequence $c_t \leq \frac{c_0}{\eta^{(1-\epsilon)}}$ for all t where $\epsilon \in (0,1)$ and c_0 is constant, and define $\overline{\mathbf{x}}_t := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_t^{(i)}$. Consider the set of synchronization indices $\mathcal{I}_T = \{I_{(1)}, I_{(2)}, \ldots, I_{(t)}, \ldots\}$. Then for any $I_{(t)} \in \mathcal{I}_T$, we have:*

$$\mathbb{E}\sum_{j=1}^{n}\left\|\overline{\mathbf{x}}^{I_{(t)}}-\mathbf{x}_j^{I_{(t)}}\right\|^2 = \mathbb{E}\|\mathbf{X}^{I_{(t)}}-\bar{\mathbf{X}}^{I_{(t)}}\|_F^2 \leq \frac{4nA\eta^2}{p^2}$$

for constant $A = \frac{p}{2}\left(2H^2G^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)\left(\frac{16}{\omega}+\frac{4}{p}\right)+\frac{2c_0\omega}{\eta^{(1-\epsilon)}}\right)$ where $p=\frac{\delta\gamma}{8}$, $\delta := 1-|\lambda_2(\mathbf{W})|$, ω is compression parameter for operator $\mathcal{C}$.

Substituting the bound from Lemma 12 in (122) and using the fact that $p \leq 1$, we have:

$$\mathbb{E}\|\mathbf{X}^{(t+1)}-\bar{\mathbf{X}}^{(t+1)}\|_F^2 \leq \frac{2\eta^2}{p}\left(2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)\left(\frac{16}{\omega}+\frac{8}{p}\right)+\frac{2c_0\omega n}{\eta^{(1-\epsilon)}}\right) \tag{123}$$

for the same constant $\epsilon > 0$ as in Lemma 12. Note that the above bound holds for all values of t.
Define $\Lambda := \frac{2}{p}\left(2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)\left(\frac{16}{\omega}+\frac{8}{p}\right)+\frac{2\omega c_0 n}{\eta^{(1-\epsilon)}}\right)$. Substituting (123) in (121) gives us:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|^2 \leq \frac{8(1-\beta)}{\eta}\frac{(f(\mathbf{x}^{(0)})-f^*)}{T}+\frac{4L\eta\bar{\sigma}^2}{n(1-\beta)}+\frac{10L^2\Lambda\eta^2}{n}+\frac{8L^2\eta^2\beta^4\bar{\sigma}^2}{n(1-\beta)^4}$$

Expanding on the value of Λ, we have:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|^2 \leq{}& \frac{8(1-\beta)}{\eta}\frac{(f(\mathbf{x}^{(0)})-f^*)}{T}+\frac{4L\eta\bar{\sigma}^2}{n(1-\beta)}\\ &+\frac{20\eta^2L^2}{pn}\left(2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)\left(\frac{16}{\omega}+\frac{8}{p}\right)\right)\\ &+\frac{40L^2\omega nc_0\eta^{(1+\epsilon)}}{pn}+\frac{8L^2\eta^2\beta^4\bar{\sigma}^2}{n(1-\beta)^4}\end{aligned}$$

Substituting the value of $\eta = (1-\beta)\sqrt{\frac{n}{T}}$, we get:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{x}}^{(t)})\right\|^2 &\leq \frac{1}{\sqrt{nT}}\left(8(f(\mathbf{x}^{(0)})-f^*)+4L\bar{\sigma}^2\right)+\frac{40L^2(1-\beta)^{(1+\epsilon)}\omega c_0 n^{(1+\epsilon)/2}}{pT^{(1+\epsilon)/2}}\\ &\quad+\frac{20(1-\beta)^2L^2}{Tp}\left(2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)\left(\frac{16}{\omega}+\frac{8}{p}\right)\right)+\frac{8L^2\beta^4\bar{\sigma}^2}{T(1-\beta)^2}\\ &\leq \frac{1}{\sqrt{nT}}\left(8(f(\mathbf{x}^{(0)})-f^*)+4L\bar{\sigma}^2\right)+\frac{40L^2\omega c_0 n^{(1+\epsilon)/2}(1-\beta)^{(1+\epsilon)}}{pT^{(1+\epsilon)/2}}\\ &\quad+\frac{80nL^2H^2G^2}{Tp}\left(\frac{16}{\omega}+\frac{8}{p}\right)+\frac{8L^2\beta^4\bar{\sigma}^2}{T(1-\beta)^2}\end{aligned}$$

where in the last inequality, we've used the fact that $(1-\beta)^r \leq 1$, $\beta^r \leq 1$ for $r > 0$. Note that we require $\eta \leq \min\left\{\frac{(1-\beta)}{4L}, \frac{(1-\beta)^2}{2\sqrt{2}L\beta^2}\right\}$, thus for $\eta = (1-\beta)\sqrt{\frac{n}{T}}$, we need to run our algorithm for $T \geq \max\left\{16L^2n, \frac{8L^2\beta^4n}{(1-\beta)^2}\right\}$ for the above rate expression to hold. We finally use the fact that $p \leq \omega$ (as $\delta \leq 1$ and $p := \frac{\gamma^*\delta}{8}$ with $\gamma^* \leq \omega$). This completes proof of the non-convex part of Theorem 2. We can further use the fact that $p \geq \frac{\delta^2\omega}{644}$ (proved in Lemma 15) to get the expression given in the theorem statement.

APPENDIX F
PROOF OF THEOREM 2 (CONVEX OBJECTIVE)

We start with the same virtual sequence defined in (15). Consider the quantity $\mathbb{E}_{\boldsymbol{\xi}^{(t)}}\|\widetilde{\mathbf{x}}^{(t+1)}-\mathbf{x}^*\|^2$, where expectation is taken over sampling across all the nodes at the t'th iteration:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi}^{(t)}}\|\widetilde{\mathbf{x}}^{(t+1)}-\mathbf{x}^*\|^2 &= \mathbb{E}_{\boldsymbol{\xi}^{(t)}}\left\|\widetilde{\mathbf{x}}^{(t)}-\frac{\eta}{(1-\beta)n}\sum_{j=1}^n\nabla F_j(\mathbf{x}_j^{(t)},\xi_j^{(t)})-\mathbf{x}^*\right\|^2\\
&= \mathbb{E}_{\boldsymbol{\xi}^{(t)}}\left\|\widetilde{\mathbf{x}}^{(t)}-\mathbf{x}^*-\frac{\eta}{(1-\beta)n}\sum_{j=1}^n\nabla f_j(\mathbf{x}_j^{(t)})+\frac{\eta}{(1-\beta)n}\sum_{j=1}^n\nabla f_j(\mathbf{x}_j^{(t)})-\frac{\eta}{n(1-\beta)}\sum_{j=1}^n\nabla F_j(\mathbf{x}_j^{(t)},\xi_j^{(t)})\right\|^2\\
&= \left\|\widetilde{\mathbf{x}}^{(t)}-\mathbf{x}^*-\frac{\eta}{(1-\beta)n}\sum_{j=1}^n\nabla f_j(\mathbf{x}_j^{(t)})\right\|^2+\frac{\eta^2}{(1-\beta)^2}\mathbb{E}_{\boldsymbol{\xi}^{(t)}}\left\|\frac{1}{n}\sum_{j=1}^n\nabla f_j(\mathbf{x}_j^{(t)})-\frac{1}{n}\sum_{j=1}^n\nabla F_j(\mathbf{x}_j^{(t)},\xi_j^{(t)})\right\|^2\\
&\quad+\frac{2\eta}{(1-\beta)n}\mathbb{E}_{\boldsymbol{\xi}^{(t)}}\left\langle\widetilde{\mathbf{x}}^{(t)}-\mathbf{x}^*-\frac{\eta}{(1-\beta)n}\sum_{j=1}^n\nabla f_j(\mathbf{x}_j^{(t)}),\sum_{j=1}^n\nabla f_j(\mathbf{x}_j^{(t)})-\sum_{j=1}^n\nabla F_j(\mathbf{x}_j^{(t)},\xi_j^{(t)})\right\rangle\\
&\le \left\|\widetilde{\mathbf{x}}^{(t)}-\mathbf{x}^*-\frac{\eta}{(1-\beta)n}\sum_{j=1}^n\nabla f_j(\mathbf{x}_j^{(t)})\right\|^2+\frac{\eta^2\bar{\sigma}^2}{(1-\beta)^2n} \qquad (124)
\end{aligned}$$

Where to get the last inequality we used the fact that $\mathbb{E}_{\xi_i^{(t)}}[\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)})]=\nabla f_i(\mathbf{x}_i^{(t)})$ for all $i\in[n]$ and the variance bound (107) from Fact 5. Now we thus consider the first term in (124):

$$\begin{aligned}
\left\|\widetilde{\mathbf{x}}^{(t)}-\mathbf{x}^*-\frac{\eta}{(1-\beta)n}\sum_{j=1}^n\nabla f_j(\mathbf{x}_j^{(t)})\right\|^2 &= \|\widetilde{\mathbf{x}}^{(t)}-\mathbf{x}^*\|^2+\frac{\eta^2}{(1-\beta)^2}\underbrace{\left\|\frac{1}{n}\sum_{j=1}^n\nabla f_j(\mathbf{x}_j^{(t)})\right\|^2}_{T_1}\\
&\quad-\frac{2\eta}{(1-\beta)}\underbrace{\left\langle\widetilde{\mathbf{x}}^{(t)}-\mathbf{x}^*,\frac{1}{n}\sum_{j=1}^n\nabla f_j(\mathbf{x}_j^{(t)})\right\rangle}_{T_2} \qquad (125)
\end{aligned}$$

To bound T_1 in (125), note that:

$$\begin{aligned}
T_1 &= \left\|\frac{1}{n}\sum_{j=1}^n(\nabla f_j(\mathbf{x}_j^{(t)})-\nabla f_j(\overline{\mathbf{x}}^{(t)})+\nabla f_j(\overline{\mathbf{x}}^{(t)})-\nabla f_j(\mathbf{x}^*))\right\|^2\\
&\le \frac{2}{n}\sum_{j=1}^n\|\nabla f_j(\mathbf{x}_j^{(t)})-\nabla f_j(\overline{\mathbf{x}}^{(t)})\|^2+2\left\|\frac{1}{n}\sum_{j=1}^n\nabla f_j(\overline{\mathbf{x}}^{(t)})-\frac{1}{n}\sum_{j=1}^n\nabla f_j(\mathbf{x}^*)\right\|^2\\
&\le \frac{2L^2}{n}\sum_{j=1}^n\|\mathbf{x}_j^{(t)}-\overline{\mathbf{x}}^{(t)}\|^2+4L(f(\overline{\mathbf{x}}^{(t)})-f^*) \qquad (126)
\end{aligned}$$

where in the last inequality, we used $L-$Lipschitz gradient property of objectives $\{f_j\}_{j=1}^n$ to bound the first term and optimality of $\mathbf{x}^*$ for f (i.e., $\nabla f(\mathbf{x}^*)=0$) and $L-$smoothness property of f to bound the second term as: $\left\|\frac{1}{n}\sum_{j=1}^n\nabla f_j(\overline{\mathbf{x}}^{(t)})-\frac{1}{n}\sum_{j=1}^n\nabla f_j(\mathbf{x}^*)\right\|^2=\left\|\nabla f(\overline{\mathbf{x}}^{(t)})-\nabla f(\mathbf{x}^*)\right\|^2\le 2L\left(f(\overline{\mathbf{x}}^{(t)})-f^*\right)$.
To bound T_2 in (125), note that:

$$\begin{aligned}
-2T_2 &= -2\left\langle\widetilde{\mathbf{x}}^{(t)}-\overline{\mathbf{x}}^{(t)},\frac{1}{n}\sum_{j=1}^n\nabla f_j(\mathbf{x}_j^{(t)})\right\rangle-\frac{2}{n}\sum_{j=1}^n\left\langle\overline{\mathbf{x}}^{(t)}-\mathbf{x}^*,\nabla f_j(\mathbf{x}_j^{(t)})\right\rangle\\
&= 2\frac{\beta^2}{(1-\beta)}\left\langle\frac{\eta}{n}\sum_{i=1}^n\mathbf{v}_i^{(t-1)},\frac{1}{n}\sum_{j=1}^n\nabla f_j(\mathbf{x}_j^{(t)})\right\rangle-\frac{2}{n}\sum_{j=1}^n\left\langle\overline{\mathbf{x}}^{(t)}-\mathbf{x}^*,\nabla f_j(\mathbf{x}_j^{(t)})\right\rangle \qquad (127)
\end{aligned}$$

In (127), we used the definition of $\widetilde{\mathbf{x}}^{(t)}$ from (13) to write $\widetilde{\mathbf{x}}^{(t)} - \overline{\mathbf{x}}^{(t)} = -\frac{\eta\beta^2}{(1-\beta)}\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i^{(t-1)}$. Now we note a simple trick for inner-products:

$$\left\langle \frac{\eta}{n}\sum_{i=1}^{n}\mathbf{v}_i^{(t-1)}, \frac{1}{n}\sum_{j=1}^{n}\nabla f_j(\mathbf{x}_j^{(t)}) \right\rangle = \left\langle \frac{(\eta)^{3/4}}{n}\sum_{i=1}^{n}\mathbf{v}_i^{(t-1)}, \frac{(\eta)^{1/4}}{n}\sum_{j=1}^{n}\nabla f_j(\mathbf{x}_j^{(t)}) \right\rangle. \tag{128}$$

This trick is crucial to getting a speedup of n – the number of worker nodes – in our final convergence rate. Using $2\langle \mathbf{a}, \mathbf{b}\rangle \leq \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2$ for bounding (128) and then substituting that in (127) gives

$$-2T_2 \leq \frac{\beta^2}{(1-\beta)}\left[(\eta)^{3/2}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i^{(t-1)}\right\|^2 + (\eta)^{1/2}\left\|\frac{1}{n}\sum_{j=1}^{n}\nabla f_j(\mathbf{x}_j^{(t)})\right\|^2\right] - \frac{2}{n}\sum_{j=1}^{n}\left\langle \overline{\mathbf{x}}^{(t)} - \mathbf{x}^*, \nabla f_j(\mathbf{x}_j^{(t)})\right\rangle \tag{129}$$

Note that the second term of (129) is the same as T_1 from (125) and we have already bounded that in (126). We now focus on bounding the last term of (129). Using expression for convexity and L-smoothness for f_j, $j \in [n]$ respectively, we can bound this as follows:

$$\begin{aligned}
-\frac{2}{n}\sum_{j=1}^{n}\langle \overline{\mathbf{x}}^{(t)} - \mathbf{x}^*, \nabla f_j(\mathbf{x}_j^{(t)})\rangle &= -\frac{2}{n}\sum_{j=1}^{n}\left[\left\langle \overline{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}, \nabla f_j(\mathbf{x}_j^{(t)})\right\rangle + \left\langle \mathbf{x}_j^{(t)} - \mathbf{x}^*, \nabla f_j(\mathbf{x}_j^{(t)})\right\rangle\right] \\
&\leq -\frac{2}{n}\sum_{j=1}^{n}\left[f_j(\overline{\mathbf{x}}^{(t)}) - f_j(\mathbf{x}_j^{(t)}) - \frac{L}{2}\|\overline{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2 + f_j(\mathbf{x}_j^{(t)}) - f_j(\mathbf{x}^*)\right] \\
&= -2(f(\overline{\mathbf{x}}^{(t)}) - f(\mathbf{x}^*)) + \frac{L}{n}\sum_{j=1}^{n}\|\overline{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2
\end{aligned} \tag{130}$$

Substituting the bounds for the second and the last terms of (129) from (126) and (130), respectively, we get

$$\begin{aligned}
-2T_2 \leq \frac{(\eta)^{3/2}\beta^2}{(1-\beta)}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i^{(t-1)}\right\|^2 + \frac{(\eta)^{1/2}\beta^2}{(1-\beta)}\left(\frac{2L^2}{n}\sum_{j=1}^{n}\|\mathbf{x}_j^{(t)} - \overline{\mathbf{x}}^{(t)}\|^2 + 4L(f(\overline{\mathbf{x}}^{(t)}) - f^*)\right) \\
- 2(f(\overline{\mathbf{x}}^{(t)}) - f(\mathbf{x}^*)) + \frac{L}{n}\sum_{j=1}^{n}\|\overline{\mathbf{x}}^{(t)} - \mathbf{x}_j^{(t)}\|^2
\end{aligned}$$

Thus we finally have:

$$\begin{aligned}
-\frac{2\eta}{(1-\beta)}T_2 \leq \frac{\eta^{5/2}\beta^2}{(1-\beta)^2}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i^{(t-1)}\right\|^2 + \left(\frac{2\eta^{3/2}\beta^2L^2}{(1-\beta)^2} + \frac{\eta L}{(1-\beta)}\right)\frac{1}{n}\sum_{j=1}^{n}\|\mathbf{x}_j^{(t)} - \overline{\mathbf{x}}^{(t)}\|^2 \\
+ \left(\frac{4\eta^{3/2}\beta^2L}{(1-\beta)^2} - \frac{2\eta}{(1-\beta)}\right)\left(f(\overline{\mathbf{x}}^{(t)}) - f^*\right)
\end{aligned} \tag{131}$$

Substituting (126), (131) in (125) and using the resulting bound back in (124), and then taking expectation w.r.t. the entire process, we get:

$$\begin{aligned}
\mathbb{E}\|\widetilde{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|^2 \leq \mathbb{E}\|\widetilde{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 + \frac{\eta^{5/2}\beta^2}{(1-\beta)^2}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i^{(t-1)}\right\|^2 + \frac{\eta^2\bar{\sigma}^2}{(1-\beta)^2 n} \\
+ \left(\frac{2\eta^2L^2}{(1-\beta)^2} + \frac{2\eta^{3/2}\beta^2L^2}{(1-\beta)^2} + \frac{\eta L}{(1-\beta)}\right)\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\|\mathbf{x}_j^{(t)} - \overline{\mathbf{x}}^{(t)}\|^2 \\
+ \left(\frac{4\eta^2L}{(1-\beta)^2} + \frac{4\eta^{3/2}\beta^2L}{(1-\beta)^2} - \frac{2\eta}{(1-\beta)}\right)\left(\mathbb{E}f(\overline{\mathbf{x}}^{(t)}) - f^*\right)
\end{aligned} \tag{132}$$

Using the fact that $\mathbb{E}\left\|\frac{1}{n}\sum_{j=1}^{n}\mathbf{v}_j^{(t)}\right\|^2 \leq \frac{G^2}{(1-\beta)^2}$ for all $t \geq 1$ (see proof of Fact 6), we have:

$$\begin{aligned}
\mathbb{E}\|\widetilde{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|^2 \leq \mathbb{E}\|\widetilde{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 + \frac{\eta^{5/2}\beta^2G^2}{(1-\beta)^4} + \frac{\eta^2\bar{\sigma}^2}{(1-\beta)^2 n} \\
+ \left(\frac{2\eta^2L^2}{(1-\beta)^2} + \frac{2\eta^{3/2}\beta^2L^2}{(1-\beta)^2} + \frac{\eta L}{(1-\beta)}\right)\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\|\mathbf{x}_j^{(t)} - \overline{\mathbf{x}}^{(t)}\|^2 \\
+ \left(\frac{4\eta^2L}{(1-\beta)^2} + \frac{4\eta^{3/2}\beta^2L}{(1-\beta)^2} - \frac{2\eta}{(1-\beta)}\right)\left(\mathbb{E}f(\overline{\mathbf{x}}^{(t)}) - f^*\right)
\end{aligned} \tag{133}$$

If we take $\eta \leq \min\left\{\frac{(1-\beta)}{8L}, \frac{(1-\beta)^2}{(8L\beta^2)^2}\right\}$, then we have:

$$\left(\frac{2\eta^2L^2}{(1-\beta)^2} + \frac{2\eta^{3/2}\beta^2L^2}{(1-\beta)^2} + \frac{\eta L}{(1-\beta)}\right) \leq \frac{3\eta L}{2(1-\beta)} \tag{134}$$

$$\left(\frac{4\eta^2L}{(1-\beta)^2} + \frac{4\eta^{3/2}\beta^2L}{(1-\beta)^2} - \frac{2\eta}{(1-\beta)}\right) \leq -\frac{\eta}{(1-\beta)} \tag{135}$$

Substituting the bounds from (134) and (135) to (133) gives

$$\mathbb{E}\|\widetilde{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|^2 \leq \mathbb{E}\|\widetilde{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 + \frac{\eta^{5/2}\beta^2G^2}{(1-\beta)^4} + \frac{\eta^2\bar{\sigma}^2}{(1-\beta)^2n} + \frac{3\eta L}{2(1-\beta)}\frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\|\mathbf{x}_j^{(t)} - \overline{\mathbf{x}}^{(t)}\|^2 - \frac{\eta}{(1-\beta)}\left(\mathbb{E}f(\overline{\mathbf{x}}^{(t)}) - f^*\right) \tag{136}$$

We can now bound the second last term in R.H.S. of (136) similar to (123) in the proof of non-convex part of Theorem 2 given in Appendix E. This gives us the bound:

$$\mathbb{E}\|\mathbf{X}^{(t+1)} - \bar{\mathbf{X}}^{(t+1)}\|_F^2 \leq \frac{2\eta^2}{p}\left(2H^2nG^2\left(1 + \frac{\beta^2}{(1-\beta)^2}\right)\left(\frac{16}{\omega} + \frac{8}{p}\right) + \frac{2c_0\omega n}{\eta^{(1-\epsilon)}}\right)$$

Using above bound for the term $\sum_{j=1}^{n}\mathbb{E}\|\mathbf{x}_j^{(t)} - \overline{\mathbf{x}}^{(t)}\|^2$ in (136) we get:

$$\mathbb{E}\|\widetilde{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|^2 \leq \mathbb{E}\|\widetilde{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 + \frac{\eta^{5/2}\beta^2G^2}{(1-\beta)^4} + \frac{\eta^2\bar{\sigma}^2}{(1-\beta)^2n} - \frac{\eta}{(1-\beta)}\left(\mathbb{E}f(\overline{\mathbf{x}}^{(t)}) - f^*\right) + \frac{3\eta^3L}{p(1-\beta)}\left(2H^2G^2\left(1 + \frac{\beta^2}{(1-\beta)^2}\right)\left(\frac{16}{\omega} + \frac{8}{p}\right) + \frac{2c_0\omega}{\eta^{(1-\epsilon)}}\right) \tag{137}$$

By rearranging terms in (137) and noting that $p \leq \omega$ (as $\delta \leq 1$ and $p := \frac{\gamma^*\delta}{8}$ with $\gamma^* \leq \omega$) and the fact that $\left(1 + \frac{\beta^2}{(1-\beta)^2}\right) \leq \frac{2}{(1-\beta)^2}$ (because $\beta < 1$), we get:

$$\mathbb{E}\|\widetilde{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|^2 \leq \mathbb{E}\|\widetilde{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 + \frac{\eta^{5/2}\beta^2G^2}{(1-\beta)^4} + \frac{\eta^2\bar{\sigma}^2}{(1-\beta)^2n} - \frac{\eta}{(1-\beta)}\left(\mathbb{E}f(\overline{\mathbf{x}}^{(t)}) - f^*\right) + \frac{288\eta^3LH^2G^2}{p^2(1-\beta)^3} + \frac{6c_0\omega L\eta^{(2+\epsilon)}}{p(1-\beta)} \tag{138}$$

Summing (138) from $t = 0$ to $T-1$, rearranging terms and diving by T both sides gives us:

$$\sum_{t=0}^{T-1}\frac{\left(\mathbb{E}f(\overline{\mathbf{x}}^{(t)}) - f^*\right)}{T} \leq \frac{(1-\beta)}{\eta}\sum_{t=0}^{T-1}\frac{\left(\mathbb{E}\|\widetilde{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 - \mathbb{E}\|\widetilde{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|^2\right)}{T} + \frac{\eta^{3/2}\beta^2G^2}{(1-\beta)^3} + \frac{\eta\bar{\sigma}^2}{(1-\beta)n} + \frac{288\eta^2LH^2G^2}{p^2(1-\beta)^2} + \frac{6c_0\omega L\eta^{(1+\epsilon)}}{p}$$

Using Jensen's inequality for convex function f on the L.H.S. and setting $\eta = (1-\beta)\sqrt{\frac{n}{T}}$ for $T \geq \max\{(8L)^2n, \frac{(8\beta^2L)^4n}{(1-\beta)^2}\}$, for $\overline{\mathbf{x}}_{avg}^{(T)} := \frac{1}{T}\sum_{t=0}^{T-1}\bar{\mathbf{x}}^{(t)}$ we have that:

$$\mathbb{E}f(\overline{\mathbf{x}}_{avg}^{(T)}) - f^* \leq \frac{\left(\mathbb{E}\|\widetilde{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2 - \mathbb{E}\|\widetilde{\mathbf{x}}^{(T)} - \mathbf{x}^*\|^2\right)}{\sqrt{nT}} + \frac{n^{3/4}\beta^2G^2}{(1-\beta)^{3/2}T^{3/4}} + \frac{\bar{\sigma}^2}{\sqrt{nT}} + \frac{288LH^2G^2}{p^2T} + \frac{6c_0\omega L(1-\beta)^{(1+\epsilon)}n^{(1+\epsilon)/2}}{pT^{(1+\epsilon)/2}}$$

Using the fact that $\widetilde{\mathbf{x}}^{(0)} = \overline{\mathbf{x}}^{(0)}$ and $\epsilon, \beta \in (0,1)$ we have:

$$\mathbb{E}f(\overline{\mathbf{x}}_{avg}^{(T)}) - f^* \leq \frac{\|\overline{\mathbf{x}}^{(0)} - \mathbf{x}^*\|^2 + \bar{\sigma}^2}{\sqrt{nT}} + \frac{n^{3/4}\beta^2G^2}{(1-\beta)^{3/2}T^{3/4}} + \frac{384nLH^2G^2}{p^2T} + \frac{6c_0\omega Ln^{(1+\epsilon)/2}}{pT^{(1+\epsilon)/2}}$$

This completes proof of convex part of Theorem 2. We can further use the fact that $p \geq \frac{\delta^2\omega}{644}$ to get the expression given in the theorem statement.

APPENDIX G
PROOF OF LEMMA 12 (CONSENSUS)

In this section, we provide a proof of Lemma 12, which states that $\sum_{j=1}^{n}\mathbb{E}\left\|\bar{\mathbf{x}}^{I_{(t)}}-\mathbf{x}_j^{I_{(t)}}\right\|^2$ – the difference between the local and the average iterates at the synchronization indices – is bounded by a constant times the learning rate η, which can effectively be made small by running the algorithm for larger number of iterations T as we choose $\eta=(1-\beta)\sqrt{\frac{n}{T}}$. Thus, this result shows that the nodes achieve a consensus towards the average parameter vector as the algorithm progresses.

We first provide a high level idea of the proof to aid the reader. Our interest is in providing a bound for $e^{(1)}_{I_{(t)}}:=\sum_{j=1}^{n}\mathbb{E}\left\|\bar{\mathbf{x}}^{I_{(t)}}-\mathbf{x}_j^{I_{(t)}}\right\|^2$. We show this by setting up a contracting recursion for $e^{(1)}_{I_{(t)}}$. First we prove that

$$e^{(1)}_{I_{(t+1)}}\leq(1-\alpha_1)e^{(1)}_{I_{(t)}}+(1-\alpha_1)e^{(2)}_{I_{(t)}}+c_1\eta^2, \tag{139}$$

where $e^{(2)}_{I_{(t)}}:=\sum_{j=1}^{n}\mathbb{E}\left\|\hat{\mathbf{x}}^{I_{(t+1)}}-\mathbf{x}_j^{I_{(t)}}\right\|^2$, $\alpha_1\in(0,1)$, and c_1 is a constant that depends on n,δ,β,H,G. The quantity $e^{(2)}_{I_{(t)}}$ relates to the expected deviation of local node parameters and their copies. Note that (139) gives a contracting recursion in $e^{(1)}_{I_{(t)}}$, but it also gives the other term $e^{(2)}_{I_{(t)}}$, which we have to bound. It turns out that we can prove a similar inequality for $e^{(2)}_{I_{(t)}}$:

$$e^{(2)}_{I_{(t+1)}}\leq(1-\alpha_2)e^{(1)}_{I_{(t)}}+(1-\alpha_2)e^{(2)}_{I_{(t)}}+c_2\eta^2, \tag{140}$$

where $\alpha_2\in(0,1)$; furthermore, we can choose α_1,α_2 such that $\alpha_1+\alpha_2>1$.

Define $e_{I_{(t)}}:=e^{(1)}_{I_{(t)}}+e^{(2)}_{I_{(t)}}$. Adding (139) and (140) gives the following recursion with $\alpha\in(0,1)$:

$$e_{I_{(t+1)}}\leq(1-\alpha)e_{I_{(t)}}+c_3\eta^2. \tag{141}$$

From (141), we can show that $e_{I_{(t)}}\leq C\eta^2$ for some C that depends on $n,\delta,\beta,H,G,\omega,c_0$. The result of Lemma 12 follows from this because $\sum_{j=1}^{n}\mathbb{E}\left\|\bar{\mathbf{x}}^{I_{(t)}}-\mathbf{x}_j^{I_{(t)}}\right\|^2=e^{(1)}_{I_{(t)}}\leq e_{I_{(t)}}$.

We first state the above-mentioned recursion results for $e^{(1)}_{I_{(t+1)}}$ and $e^{(2)}_{I_{(t+1)}}$ below in Lemma 13 and Lemma 14, respectively, and then using that we prove Lemma 12. The proofs of Lemma 13 and Lemma 14 are provided in Appendix H.

Lemma 13. *Under the setting of Theorem 2,* $e^{(1)}_{I_{(t+1)}}:=\sum_{j=1}^{n}\mathbb{E}\left\|\bar{\mathbf{x}}^{I_{(t+1)}}-\mathbf{x}_j^{I_{(t+1)}}\right\|^2$ *satisfies:*

$$e^{(1)}_{I_{(t+1)}}\leq(1+\alpha_5^{-1})R_1e^{(1)}_{I_{(t)}}+(1+\alpha_5^{-1})R_2e^{(2)}_{I_{(t)}}+Q_1\eta^2,$$

where $R_1=(1+\alpha_1)(1-\gamma\delta)^2, R_2=(1+\alpha_1^{-1})\gamma^2\lambda^2$ *and* $Q_1=2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)(1+\alpha_5)(R_1+R_2)$. *Here* $\alpha_1,\alpha_5>0$, δ *is the spectral gap,* H *is the synchronization gap,* γ *is the consensus stepsize, and* $\lambda:=\|\mathbf{W}-\mathbf{I}\|_2$ *where* $\mathbf{W}$ *is a doubly stochastic mixing matrix.*

Lemma 14. *Under the setting of Theorem 2,* $e^{(2)}_{I_{(t+1)}}:=\sum_{j=1}^{n}\mathbb{E}\left\|\hat{\mathbf{x}}^{I_{(t+2)}}-\mathbf{x}_j^{I_{(t+1)}}\right\|^2$ *satisfies:*

$$e^{(2)}_{I_{(t+1)}}\leq(1+\alpha_5^{-1})R_3e^{(2)}_{I_{(t)}}+(1+\alpha_5^{-1})R_4e^{(1)}_{I_{(t)}}+\eta^2Q_2,$$

where $R_3=(1+\gamma\lambda)^2(1+\alpha_4)(1+\alpha_3)(1+\alpha_2)(1-\omega)$, $R_4=\gamma^2\lambda^2(1+\alpha_4^{-1})(1+\alpha_3)(1+\alpha_2)(1-\omega)$ *and* $Q_2=2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)((1+\alpha_5)(R_3+R_4)+(1+\alpha_2^{-1})+(1+\alpha_3^{-1})(1+\alpha_2)(1-\omega))+(1+\alpha_2)\omega n\frac{c_0}{\eta^{(1-\epsilon)}}$. *Note that* Q_2 *depends on* t *(as captured by* $c_{I_{(t)}}$ *in the expression) as we allow for our triggering threshold to change with time. Here* $\alpha_2,\alpha_3,\alpha_4>0,\alpha_5>0$ *are the same as those used in Lemma 13,* δ *is the spectral gap,* H *is the synchronization gap,* γ *is the consensus stepsize, and* $\lambda=\|\mathbf{W}-\mathbf{I}\|_2$ *where* $\mathbf{W}$ *is a doubly stochastic mixing matrix.*

Proof of Lemma 12. Having established the bounds on $e^{(1)}_{I_{(t+1)}}$ and $e^{(2)}_{I_{(t+1)}}$, we are now ready to prove Lemma 12. Consider the following expression:

$$e_{I_{(t+1)}}=\underbrace{\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2}_{e^{(1)}_{I_{(t+1)}}}+\underbrace{\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2}_{e^{(2)}_{I_{(t+1)}}} \tag{142}$$

We note that Lemma 13 and Lemma 14 provide bounds for the first and the second term in the RHS of (142). Substituting them in (142) gives:

$$e_{I_{(t+1)}}\leq R_1(1+\alpha_5^{-1})\mathbb{E}\left\|\bar{\mathbf{X}}^{I_{(t)}}-\mathbf{X}^{I_{(t)}}\right\|^2+R_2(1+\alpha_5^{-1})\mathbb{E}\left\|\hat{\mathbf{X}}^{I_{(t+1)}}-\mathbf{X}^{I_{(t)}}\right\|^2$$

$$+ R_4(1+\alpha_5^{-1})\mathbb{E}\left\|\bar{\mathbf{X}}^{I_{(t)}} - \mathbf{X}^{I_{(t)}}\right\|^2 + R_3(1+\alpha_5^{-1})\mathbb{E}\left\|\hat{\mathbf{X}}^{I_{(t+1)}} - \mathbf{X}^{I_{(t)}}\right\|^2 + (Q_1+Q_2)\eta^2 \tag{143}$$

Define the following:

$$\pi_1(\gamma) := R_2 + R_3 = \gamma^2\lambda^2(1+\alpha_1^{-1}) + (1+\gamma\lambda)^2(1+\alpha_4)(1+\alpha_3)(1+\alpha_2)(1-\omega) \tag{144}$$

$$\pi_2(\gamma) := R_1 + R_4 = (1-\delta\gamma)^2(1+\alpha_1) + \gamma^2\lambda^2(1+\alpha_4^{-1})(1+\alpha_3)(1+\alpha_2)(1-\omega) \tag{145}$$

$$\begin{aligned}\pi_0 :=& Q_1 + Q_2 \leq 2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)(1+\alpha_5)(R_1+R_2+R_3+R_4)\\ &+2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)((1+\alpha_2^{-1}) + (1-\omega)(1+\alpha_3^{-1})(1+\alpha_2)) + (1+\alpha_2)\frac{\omega n c_0}{\eta^{(1-\epsilon)}}\end{aligned} \tag{146}$$

The bound on $e_{I_{(t+1)}}$ in (143) can be rewritten as:

$$\begin{aligned}e_{I_{(t+1)}} &\leq (1+\alpha_5^{-1})\left[\pi_1(\gamma)\mathbb{E}\|\mathbf{X}^{I_{(t)}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 + \pi_2(\gamma)\mathbb{E}\|\mathbf{X}^{I_{(t)}} - \bar{\mathbf{X}}^{I_{(t)}}\|_F^2\right] + \pi_0\eta^2\\ &\leq (1+\alpha_5^{-1})\max\{\pi_1(\gamma),\pi_2(\gamma)\}\,\mathbb{E}\left[\|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 + \|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}}\|_F^2\right] + \pi_0\eta^2\end{aligned} \tag{147}$$

Calculation of $\max\{\pi_1(\gamma),\pi_2(\gamma)\}$ and π_0 is given in Lemma 15 in Appendix G-A, where we show that: $\max\{\pi_1(\gamma),\pi_2(\gamma)\} \leq (1-p)$ and $\pi_0 \leq \left(2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)\left(\frac{16}{\omega}+\frac{4}{p}\right) + 2\omega n\frac{c_0}{\eta^{(1-\epsilon)}}\right)$, where $p := \frac{\gamma^*\delta}{8}$. Here $\gamma^* = \frac{2\delta\omega}{64\delta+\delta^2+16\lambda^2+8\delta\lambda^2-16\delta\omega}$ is the consensus step-size. Substituting these bounds and $\alpha_5 = \frac{2}{p}$ in (147) gives:

$$\begin{aligned}e_{I_{(t+1)}} \leq& (1+\frac{p}{2})(1-p)\,\mathbb{E}\left[\|\mathbf{X}^{I_{(t)}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 + \|\mathbf{X}^{I_{(t)}} - \bar{\mathbf{X}}^{I_{(t)}}\|_F^2\right]\\ &+\left(2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)\left(\frac{16}{\omega}+\frac{4}{p}\right) + 2\omega n\frac{c_0}{\eta^{(1-\epsilon)}}\right)\eta^2.\end{aligned} \tag{148}$$

Note that $e_{I_{(t)}} = \mathbb{E}\left[\|\mathbf{X}^{I_{(t)}} - \bar{\mathbf{X}}^{I_{(t)}}\|_F^2 + \|\mathbf{X}^{I_{(t)}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2\right]$. We can write (148) as a recurrence relation for $e_{I_{(t)}}$ as:

$$e_{I_{(t+1)}} \leq \left(1-\frac{p}{2}\right)e_{I_{(t)}} + \frac{2nA}{p}\eta^2. \tag{149}$$

where $A := \frac{p}{2n}\left(2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)\left(\frac{16}{\omega}+\frac{4}{p}\right) + 2\omega n\frac{c_0}{\eta^{(1-\epsilon)}}\right)$. Using (149), it can be shown (proved in Lemma 16 in Appendix G-A below) that for all $I_{(t)} \in \mathcal{I}_T$, we have:

$$e_{I_{(t)}} \leq \frac{4nA\eta^2}{p^2}$$

Note that we also have: $\mathbb{E}\|\bar{\mathbf{X}}^{I_{(t)}} - \mathbf{X}^{I_{(t)}}\|_F^2 \leq \mathbb{E}\left[\|\bar{\mathbf{X}}^{I_{(t)}} - \mathbf{X}^{I_{(t)}}\|_F^2 + \|\hat{\mathbf{X}}^{I_{(t+1)}} - \mathbf{X}^{I_{(t)}}\|_F^2\right] = e_{I_{(t)}}$. Thus, we get the following result for any synchronization index $I_{(t)} \in \mathcal{I}_T$:

$$\mathbb{E}\|\bar{\mathbf{X}}^{I_{(t)}} - \mathbf{X}^{I_{(t)}}\|_F^2 \leq \frac{4nA\eta^2}{p^2},$$

where $A = \frac{p}{2}\left(2H^2G^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)\left(\frac{16}{\omega}+\frac{4}{p}\right) + 2\omega\frac{c_0}{\eta^{1-\epsilon}}\right)$ for $p = \frac{\delta\gamma^*}{8}$, $\epsilon > 0$ and $\gamma^* = \frac{2\delta\omega}{64\delta+\delta^2+16\beta^2+8\delta\beta^2-16\delta\omega}$ is the chosen consensus step size. This completes the proof for Lemma 12

□

A. Supporting Lemmas for Proving Lemma 12

Lemma 15. *Consider the following variables:*

$$\begin{aligned}\pi_1(\gamma) &:= \gamma^2\lambda^2(1+\alpha_1^{-1}) + (1+\gamma\lambda)^2(1+\alpha_4)(1+\alpha_3)(1+\alpha_2)(1-\omega)\\ \pi_2(\gamma) &:= (1-\delta\gamma)^2(1+\alpha_1) + \gamma^2\lambda^2(1+\alpha_4^{-1})(1+\alpha_3)(1+\alpha_2)(1-\omega)\\ \pi_0 &:= 2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)(1+\alpha_5)(\pi_1(\gamma)+\pi_2(\gamma))\\ &\quad + 2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)((1+\alpha_2^{-1}) + (1-\omega)(1+\alpha_3^{-1})(1+\alpha_2)) + (1+\alpha_2)\omega n\frac{c_0}{\eta^{(1-\epsilon)}}\end{aligned}$$

and the following choice of variables:

$$\alpha_1 := \frac{\gamma\delta}{2},\ \alpha_2 := \frac{\omega}{4},\ \alpha_3 := \frac{\omega}{4},\ \alpha_4 := \frac{\omega}{4},\ \alpha_5 := \frac{2}{p}$$

$$p := \frac{\delta\gamma^*}{8},\ \gamma^* := \frac{2\delta\omega}{64\delta + \delta^2 + 16\lambda^2 + 8\delta\lambda^2 - 16\delta\omega}$$

Then, it can be shown that:

$$\max\{\pi_1(\gamma^*), \pi_2(\gamma^*)\} \leq 1 - \frac{\delta^2\omega}{644} \quad , \quad \pi_0 \leq 2H^2nG^2\left(1 + \frac{\beta^2}{(1-\beta)^2}\right)\left(\frac{16}{\omega} + \frac{4}{p}\right) + 2\omega n\frac{c_0}{\eta^{(1-\epsilon)}}$$

Proof. We adapt a part of the proof of [Theorem 1] [19] to prove Lemma 15. Consider:

$$\begin{aligned}(1+\alpha_4)(1+\alpha_3)(1+\alpha_2)(1-\omega) &= (1+\frac{\omega}{4})^3(1-\omega)\\ &= \left(1 - \frac{\omega^4}{64} - \frac{11\omega^3}{64} - \frac{9\omega^2}{16} - \frac{\omega}{4}\right)\\ &\leq \left(1 - \frac{\omega}{4}\right)\end{aligned}$$

This gives us:

$$\pi_1(\gamma) \leq \gamma^2\lambda^2\left(1 + \frac{2}{\gamma\delta}\right) + (1+\gamma\lambda)^2\left(1 - \frac{\omega}{4}\right)$$

Noting that $\gamma^2 \leq \gamma$ (for $\gamma \leq 1$ which is true for γ^*) and $\lambda \leq 2$, we have:

$$\pi_1(\gamma) \leq \lambda^2\left(\gamma + \frac{2\gamma}{\delta}\right) + (1+8\gamma)\left(1 - \frac{\omega}{4}\right)$$

Substituting value of γ^* in above, it can be shown that:

$$\pi_1(\gamma^*) \leq 1 - \frac{\delta^2\omega}{4(64\delta + \delta^2 + 16\lambda^2 + 8\delta\lambda^2 - 16\delta\omega)}$$

Now we note that:

$$\pi_2(\gamma) = (1-\delta\gamma)^2\left(1 + \frac{\delta\gamma}{2}\right) + \gamma^2\lambda^2\left(1 + \frac{4}{\omega}\right)\left(1 + \frac{\omega}{4}\right)^2(1-\omega)$$

Noting the fact that for $x = \delta\gamma \leq 1$, we have $(1-x)^2\left(1+\frac{x}{2}\right) \leq (1-x)\left(1-\frac{x}{2}\right)$,

$$\begin{aligned}\pi_2(\gamma) &\leq \left(1 - \frac{\gamma\delta}{2}\right)^2 + \gamma^2\lambda^2\left(1 + \frac{4}{\omega}\right)\left(1 + \frac{\omega}{4}\right)^2(1-\omega)\\ &= \left(1 - \frac{\gamma\delta}{2}\right)^2 + \gamma^2\lambda^2\left(3 + \frac{3\omega}{4} + \frac{\omega^2}{16} + \frac{4}{\omega}\right)(1-\omega)\\ &\leq \left(1 - \frac{\gamma\delta}{2}\right)^2 + \gamma^2\lambda^2\frac{4}{\omega} =: \zeta(\gamma)\end{aligned}$$

Note that $\zeta(\gamma)$ is convex and quadratic in γ, and attains minima at $\gamma' = \frac{2\delta\omega}{16\lambda^2+\delta^2\omega}$ with value $\zeta(\gamma') = \frac{16\lambda^2}{16\lambda^2+\omega\delta^2}$.

By the Jensen's inequality, we note that for any $s \in [0,1]$

$$\zeta(s\gamma') \leq (1-s)\zeta(0) + s\zeta(\gamma') = 1 - s\frac{\delta^2\omega}{16\lambda^2 + \delta^2\omega}$$

For the choice $s = \frac{16\lambda^2+\omega\delta^2}{64\delta+\delta^2+16\lambda^2+8\delta\lambda^2-16\delta\omega}$, it can be seen that $s\gamma' = \gamma^*$. Thus we get:

$$\begin{aligned}\pi_2(\gamma^*) \leq \zeta(s\gamma') &\leq 1 - \frac{\delta^2\omega}{(64\delta + \delta^2 + 16\lambda^2 + 8\delta\lambda^2 - 16\delta\omega)}\\ &\leq 1 - \frac{\delta^2\omega}{4(64\delta + \delta^2 + 16\lambda^2 + 8\delta\lambda^2 - 16\delta\omega)}\end{aligned}$$

Thus we have:

$$\max\{\pi_1(\gamma^*), \pi_2(\gamma^*)\} \leq 1 - \frac{\delta^2\omega}{4(64\delta + \delta^2 + 16\lambda^2 + 8\delta\lambda^2 - 16\delta\omega)}.$$

Using the value of γ^* given in the lemma statement, we have $\frac{\delta^2\omega}{4(64\delta+\delta^2+16\lambda^2+8\delta\lambda^2-16\delta\omega)} = \frac{\delta\gamma^*}{8}$. Define $p := \frac{\gamma^*\delta}{8}$. Using crude estimates $\delta \leq 1, \omega \geq 0, \lambda \leq 2$, we can lower-bound p as $p \geq \frac{\delta^2\omega}{644}$. Thus we have

$$\max\{\pi_1(\gamma^*), \pi_2(\gamma^*)\} \leq 1 - \frac{\delta^2\omega}{644}.$$

Now we upper-bound the value of π_0:

$$\begin{aligned}
\pi_0 &:= 2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)(1+\alpha_5)(\pi_1(\gamma)+\pi_2(\gamma)) + (1+\alpha_2)\omega n\frac{c_0}{\eta^{(1-\epsilon)}} \\
&\quad + 2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)((1+\alpha_2^{-1}) + (1-\omega)(1+\alpha_3^{-1})(1+\alpha_2)) \\
&\leq 4H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)(1+\frac{2}{p})(1-p) + (1+\frac{\omega}{4})\omega n\frac{c_0}{\eta^{(1-\epsilon)}} \\
&\quad + 2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)((1+\frac{4}{\omega}) + (1-\omega)(1+\frac{4}{\omega})(1+\frac{\omega}{4})) \\
&\leq 4H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)\frac{2}{p} + (1+\frac{\omega}{4})\omega n\frac{c_0}{\eta^{(1-\epsilon)}} + 2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)(1+\frac{8}{\omega})
\end{aligned}$$

Where in the first inequality we have used the fact that $\pi_1(\gamma)+\pi_2(\gamma) \leq 2(1-p)$. In the second inequality, we use the fact that $(1+\frac{2}{p})(1-p) \leq \frac{2}{p}$ and $(1-\omega)(1+\frac{4}{\omega})(1+\frac{\omega}{4}) \leq \frac{4}{\omega}$. Noting that for $\omega \leq 1$, we have $(1+\frac{\omega}{4}) \leq 2$ and $\left(1+\frac{8}{\omega}\right) \leq \frac{16}{\omega}$. Using these, we have:

$$\pi_0 \leq 2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)\left(\frac{4}{p}+\frac{16}{\omega}\right) + 2\omega nHc_t.$$

This completes the proof of Lemma 15. □

Lemma 16. *Consider the sequence $\{e_{I_{(t)}}\}$ given by*

$$e_{I_{(t+1)}} \leq \left(1-\frac{p}{2}\right)e_{I_{(t)}} + \frac{2nA}{p}\eta^2,$$

where $\mathcal{I}_T = \{I_{(1)}, I_{(2)}, \ldots, I_{(t)}, \ldots\} \in [T]$ denotes the set of synchronization indices. For a parameter $p > 0$, positive constants A and η , we have:

$$e_{I_{(t)}} \leq \frac{4nA}{p^2}\eta^2$$

Proof. The proof uses an induction argument. Note that the base case is satisfied as $e_0 = 0$. Assuming the bound holds for $e_{I_{(t)}}$, for $e_{I_{(t+1)}}$, we have:

$$\begin{aligned}
e_{I_{(t+1)}} &\leq (1-\frac{p}{2})\frac{4nA\eta^2}{p^2} + \frac{2nA\eta^2}{p} \\
&= \frac{4nA\eta^2}{p^2}
\end{aligned}$$

Thus $e_{I_{(t)}} \leq \frac{4nA}{p^2}\eta^2$ for all $I_{(t)} \in \mathcal{I}_T$ from induction argument, which completes the proof. □

APPENDIX H
SUPPORTING LEMMAS FOR PROOF OF LEMMA 12

As discussed in Appendix G, the proof for Lemma 12 relies on establishing a recurrence relation between two quantities of interest: $e^{(1)}_{I_{(t)}} := \sum_{j=1}^n \mathbb{E}\left\|\bar{\mathbf{x}}^{I_{(t)}} - \mathbf{x}_j^{I_{(t)}}\right\|^2$ – the average deviation of local parameter copies and the global parameter – and $e^{(2)}_{I_{(t)}} := \sum_{j=1}^n \mathbb{E}\left\|\hat{\mathbf{x}}^{I_{(t+1)}} - \mathbf{x}_j^{I_{(t)}}\right\|^2$ – the average deviation of the local parameter and their copies. In this section, we provide a recursion relation for both $e^{(1)}_{I_{(t+1)}}$ and $e^{(2)}_{I_{(t+1)}}$, each in terms of $e^{(1)}_{I_{(t)}}$ and $e^{(2)}_{I_{(t)}}$. These results are stated in Lemma 13 and 14, respectively, which we prove below. In order to prove these lemmas we use some techniques from proof of Lemma 1 and Lemma 2 in [19].

In matrix notation, these quantities are given by:

$$\begin{aligned}
e^{(1)}_{I_{(t+1)}} &= \mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2 \\
e^{(2)}_{I_{(t+1)}} &= \mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2
\end{aligned}$$

A. Proof of Lemma 13

Using the update equations of $\mathbf{X}^{I_{(t+1)}}$ in matrix form given in (5)-(8) in Section IV, we have:

$$\|\mathbf{X}^{I_{(t+1)}} - \bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2 = \|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+1)}} + \gamma\hat{\mathbf{X}}^{I_{(t+1)}}(\mathbf{W}-\mathbf{I})\|_F^2$$

Noting that $\bar{\mathbf{X}}^{I_{(t+1)}} = \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}}$ (from (10)) and $\bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}}(\mathbf{W}-\mathbf{I}) = 0$ (from (9)), we get:

$$\|\mathbf{X}^{I_{(t+1)}} - \bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2 = \|(\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}})((1-\gamma)\mathbf{I}+\gamma\mathbf{W}) + \gamma(\hat{\mathbf{X}}^{I_{(t+1)}} - \mathbf{X}^{I_{(t+\frac{1}{2})}})(\mathbf{W}-\mathbf{I})\|_F^2$$

For any positive constant[11] α_1, we have:

$$\begin{aligned}\|\mathbf{X}^{I_{(t+1)}} - \bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2 \leq (1+\alpha_1)\|(\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}})((1-\gamma)\mathbf{I}+\gamma\mathbf{W})\|_F^2 \\ + (1+\alpha_1^{-1})\|\gamma(\hat{\mathbf{X}}^{I_{(t+1)}} - \mathbf{X}^{I_{(t+\frac{1}{2})}})(\mathbf{W}-\mathbf{I})\|_F^2\end{aligned}$$

Using $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F\|\mathbf{B}\|_2$ for any matrices $\mathbf{A},\mathbf{B}$, we have:

$$\begin{aligned}\|\mathbf{X}^{I_{(t+1)}} - \bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2 \leq (1+\alpha_1)\|(\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}})((1-\gamma)\mathbf{I}+\gamma\mathbf{W})\|_F^2 \\ + (1+\alpha_1^{-1})\gamma^2\|(\hat{\mathbf{X}}^{I_{(t+1)}} - \mathbf{X}^{I_{(t+\frac{1}{2})}})\|_F^2.\|(\mathbf{W}-\mathbf{I})\|_2^2 \qquad (150)\end{aligned}$$

To bound the first term in (150), we use the triangle inequality for Frobenius norm, giving us:

$$\|(\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}})((1-\gamma)\mathbf{I}+\gamma\mathbf{W})\|_F \leq (1-\gamma)\|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}}\|_F + \gamma\|(\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}})\mathbf{W}\|_F$$

Since $\left(\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}}\right)\frac{\mathbf{1}\mathbf{1}^T}{n} = \mathbf{0}$ (from (9)), adding this inside the last term above, we get:

$$\begin{aligned}\|(\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}})((1-\gamma)\mathbf{I}+\gamma\mathbf{W})\|_F \leq (1-\gamma)\|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}}\|_F \\ + \gamma\left\|(\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}})\left(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\right\|_F\end{aligned}$$

Using $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F\|\mathbf{B}\|_2$ and then using (112) from Fact 3 with $k=1$, we can simplify the above to:

$$\|(\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}})((1-\gamma)\mathbf{I}+\gamma\mathbf{W})\|_F \leq (1-\gamma\delta)\|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}}\|_F$$

Substituting the above in (150) and using $\lambda = \max_i\{1-\lambda_i(\mathbf{W})\} \Rightarrow \|\mathbf{W}-\mathbf{I}\|_2^2 \leq \lambda^2$, we get:

$$\|\mathbf{X}^{I_{(t+1)}} - \bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2 \leq (1+\alpha_1)(1-\gamma\delta)^2\|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}}\|_F^2 + (1+\alpha_1^{-1})\gamma^2\lambda^2\|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2$$

Taking expectation w.r.t. the entire process, we have:

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2 \leq (1+\alpha_1)(1-\gamma\delta)^2\mathbb{E}\|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}}\|_F^2 + (1+\alpha_1^{-1})\gamma^2\lambda^2\mathbb{E}\|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2$$

Define $R_1 = (1+\alpha_1)(1-\gamma\delta)^2, R_2 = (1+\alpha_1^{-1})\gamma^2\lambda^2$. Using the update steps of algorithm given in equations (6) and (10) (given in Section IV), we have:

$$\begin{aligned}\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2 \leq R_1\mathbb{E}\left\|\bar{\mathbf{X}}^{I_{(t)}} - \mathbf{X}^{I_{(t)}} - \sum_{t'=I_{(t)}}^{I_{(t+1)}-1}\eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')}))\left(\frac{\mathbf{1}\mathbf{1}^T}{n} - I\right)\right\|_F^2 \\ + R_2\mathbb{E}\left\|\hat{\mathbf{X}}^{I_{(t+1)}} - \mathbf{X}^{I_{(t)}} + \sum_{t'=I_{(t)}}^{I_{(t+1)}-1}\eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')}))\right\|_F^2\end{aligned}$$

Thus, for any $\alpha_5 > 0$ (using Footnote 11), we have:

$$\begin{aligned}\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2 \leq R_1(1+\alpha_5^{-1})\mathbb{E}\left\|\bar{\mathbf{X}}^{I_{(t)}} - \mathbf{X}^{I_{(t)}}\right\|^2 + R_2(1+\alpha_5^{-1})\mathbb{E}\left\|\hat{\mathbf{X}}^{I_{(t+1)}} - \mathbf{X}^{I_{(t)}}\right\|^2 \\ + R_1(1+\alpha_5)\mathbb{E}\left\|\sum_{t'=I_{(t)}}^{I_{(t+1)}-1}\eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')}))\left(\frac{\mathbf{1}\mathbf{1}^T}{n} - I\right)\right\|_F^2\end{aligned}$$

[11]For any two matrices $\mathbf{A},\mathbf{B} \in \mathbb{R}^{p\times q}$ and for any $\alpha > 0$, we have the following relationship for the Frobenius norm:

$$\|\mathbf{A}+\mathbf{B}\|_F^2 \leq (1+\alpha)\|\mathbf{A}\|_F^2 + (1+\alpha^{-1})\|\mathbf{B}\|_F^2$$

$$+ R_2(1+\alpha_5)\mathbb{E}\left\|\sum_{t'=I_{(t)}}^{I_{(t+1)}-1}\eta(\beta\mathbf{V}^{(t')}+\boldsymbol{\nabla}\boldsymbol{F}(\mathbf{X}^{(t')},\boldsymbol{\xi}^{(t')}))\right\|_F^2$$

Using $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_2$ to split the third term, and then using the bound $\left\|\frac{\mathbf{1}\mathbf{1}^T}{n}-\mathbf{I}\right\|_2 = 1$ (which is shown in Claim 2 in Appendix D), and further using the bound in (108) for the third and the fourth terms, the above can be rewritten as:

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\bar{\mathbf{X}}^{I_{(t+1)}}\|_F^2 \leq R_1(1+\alpha_5^{-1})\mathbb{E}\left\|\bar{\mathbf{X}}^{I_{(t)}}-\mathbf{X}^{I_{(t)}}\right\|^2 + R_2(1+\alpha_5^{-1})\mathbb{E}\left\|\hat{\mathbf{X}}^{I_{(t+1)}}-\mathbf{X}^{I_{(t)}}\right\|^2$$
$$+ 2\eta^2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)(1+\alpha_5)(R_1+R_2)$$

Defining $Q_1 = 2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)(1+\alpha_5)(R_1+R_2)$ completes the proof of Lemma 13.

B. Proof of Lemma 14

Since $\hat{\mathbf{X}}^{I_{(t+2)}} = \hat{\mathbf{X}}^{I_{(t+1)}} + \mathcal{C}((\mathbf{X}^{I_{(t+\frac{3}{2})}} - \hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)})$ (from (7) in Section IV), we have:

$$e^{(2)}_{I_{(t+1)}} = \mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 = \mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\hat{\mathbf{X}}^{I_{(t+1)}}-\mathcal{C}((\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2$$
$$= \mathbb{E}\|\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}}+\mathbf{X}^{I_{(t+1)}}-\mathbf{X}^{I_{(t+\frac{3}{2})}}-\mathcal{C}((\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2$$

For any $\alpha_2 > 0$, using result from Footnote 11, we have:

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \leq (1+\alpha_2)\mathbb{E}\|\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}}-\mathcal{C}((\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2$$
$$+ (1+\alpha_2^{-1})\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\mathbf{X}^{I_{(t+\frac{3}{2})}}\|_F^2 \tag{151}$$

The last term in R.H.S. of (151) can be bounded by using the update step (6) and then using (108) from Fact 6, which gives:

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\mathbf{X}^{I_{(t+\frac{3}{2})}}\|_F^2 \leq 2\eta^2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right) \tag{152}$$

Using the bound (152) in (151), we get:

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \leq (1+\alpha_2)\mathbb{E}\|\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}}-\mathcal{C}((\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2$$
$$+ (1+\alpha_2^{-1})2\eta^2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)$$

Note that both $\mathbf{P}^{(I_{(t+2)}-1)}$ and $\mathbf{I}-\mathbf{P}^{(I_{(t+2)}-1)}$ are diagonal matrices, with disjoint support on the diagonal entries, which implies that $\mathbb{E}\|\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 = \mathbb{E}\|(\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)}\|_F^2 + \mathbb{E}\|(\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}})(\mathbf{I}-\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2$. We get:

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \leq (1+\alpha_2)\mathbb{E}\|(\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)}-\mathcal{C}((\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2$$
$$+ (1+\alpha_2)\mathbb{E}\|(\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}})(\mathbf{I}-\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2 + 2(1+\alpha_2^{-1})\eta^2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)$$

Using the compression property (2) of operator $\mathcal{C}$, we have:

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \leq (1+\alpha_2)(1-\omega)\mathbb{E}\|(\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}})\mathbf{P}^{(I_{(t+2)}-1)}\|_F^2$$
$$+ (1+\alpha_2)\mathbb{E}\|(\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}})(\mathbf{I}-\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2 + 2(1+\alpha_2^{-1})\eta^2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)$$

Adding and subtracting $(1+\alpha_2)(1-\omega)\mathbb{E}\|(\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}})(\mathbf{I}-\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2$, we get:

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \leq (1+\alpha_2)(1-\omega)\mathbb{E}\|\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 + (1+\alpha_2^{-1})2\eta^2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)$$
$$+ (1+\alpha_2)\omega\mathbb{E}\|(\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}})(\mathbf{I}-\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2$$

To bound the third term in the RHS above, note that $\hat{\mathbf{X}}^{I_{(t+2)}-1} = \hat{\mathbf{X}}^{I_{(t+1)}}$, because $\hat{\mathbf{X}}$ does not change in between the synchronization indices, which implies that $\mathbb{E}\|(\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}})(\mathbf{I}-\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2 = \mathbb{E}\|(\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+2)}-1})(\mathbf{I}-\mathbf{P}^{(I_{(t+2)}-1)})\|_F^2$, which we can upper-bound using (111) by $nc_{I_{(t+2)}-1}\eta^2$. Using $c_t \leq \frac{c_0}{\eta^{(1-\epsilon)}}$ for all t, we get:

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}}-\hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \leq (1+\alpha_2)(1-\omega)\mathbb{E}\|\mathbf{X}^{I_{(t+\frac{3}{2})}}-\hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 + (1+\alpha_2)\omega nc_0\eta^{(1+\epsilon)}$$

$$+ (1+\alpha_2^{-1})2\eta^2 H^2 nG^2 \left(1 + \frac{\beta^2}{(1-\beta)^2}\right) \tag{153}$$

We now bound the first term in the R.H.S. of (153). From the update equation (6), we have:

$$\mathbb{E}\|\mathbf{X}^{I_{(t+\frac{3}{2})}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 = \mathbb{E}\left\|\mathbf{X}^{I_{(t+1)}} - \sum_{t'=I_{(t+1)}}^{I_{(t+2)}-1} \eta(\beta \mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')})) - \hat{\mathbf{X}}^{I_{(t+1)}}\right\|_F^2$$
$$\leq (1+\alpha_3)\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 + (1+\alpha_3^{-1})2\eta^2 H^2 nG^2 \left(1 + \frac{\beta^2}{(1-\beta)^2}\right) \tag{154}$$

where for the last inequality, α_3 is any positive constant (from Footnote 11) and we have used (108) from Fact 6. Substituting the bound (154) in (153), we have:

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \leq (1+\alpha_3)(1+\alpha_2)(1-\omega)\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2$$
$$+ (1+\alpha_3^{-1})(1+\alpha_2)(1-\omega)2\eta^2 H^2 nG^2 \left(1 + \frac{\beta^2}{(1-\beta)^2}\right)$$
$$+ (1+\alpha_2)\omega n c_0 \eta^{(1+\epsilon)} + (1+\alpha_2^{-1})2\eta^2 H^2 nG^2 \left(1 + \frac{\beta^2}{(1-\beta)^2}\right) \tag{155}$$

We now bound the first term in R.H.S. of (155). From the update equation (8) and using the fact that $\bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}}(\mathbf{W} - \mathbf{I}) = 0$, we have:

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 = \mathbb{E}\|(\mathbf{X}^{I_{(t+\frac{1}{2})}} - \hat{\mathbf{X}}^{I_{(t+1)}})((1+\gamma)\mathbf{I} - \gamma\mathbf{W}) + \gamma(\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}})(\mathbf{W} - \mathbf{I})\|_F^2$$
$$\leq (1+\alpha_4)(1+\gamma\lambda)^2\mathbb{E}\|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 + \gamma^2\lambda^2(1+\alpha_4^{-1})\mathbb{E}\|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}}\|_F^2 \tag{156}$$

where α_4 is any positive constant (from Footnote 11) and the fact that $\|(1+\gamma)\mathbf{I} - \gamma\mathbf{W}\|_2 = \|I + \gamma(\mathbf{I} - \mathbf{W})\|_2 = 1 + \gamma\|\mathbf{I} - \mathbf{W}\|_2 = 1 + \gamma\lambda$ (by definition of $\lambda = \max_i\{1 - \lambda_i(\mathbf{W})\}$) and $\|\mathbf{I} - \mathbf{W}\|_2 = \lambda$ along with $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_2$. Using the bound from (156) in (155), we get:

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \leq (1+\gamma\lambda)^2(1+\alpha_4)(1+\alpha_3)(1+\alpha_2)(1-\omega)\mathbb{E}\|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2$$
$$+ \gamma^2\lambda^2(1+\alpha_4^{-1})(1+\alpha_3)(1+\alpha_2)(1-\omega)\mathbb{E}\|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}}\|_F^2$$
$$+ 2\left((1+\alpha_2^{-1}) + (1+\alpha_3^{-1})(1+\alpha_2)(1-\omega)\right)\eta^2 H^2 nG^2 \left(1 + \frac{\beta^2}{(1-\beta)^2}\right)$$
$$+ (1+\alpha_2)\omega n c_0 \eta^{(1+\epsilon)}$$

Define $R_3 = (1+\gamma\lambda)^2(1+\alpha_4)(1+\alpha_3)(1+\alpha_2)(1-\omega)$, $R_4 = \gamma^2\lambda^2(1+\alpha_4^{-1})(1+\alpha_3)(1+\alpha_2)(1-\omega)$ and $R_5 = 2\left((1+\alpha_2^{-1}) + (1+\alpha_3^{-1})(1+\alpha_2)(1-\omega)\right)H^2 nG^2\left(1 + \frac{\beta^2}{(1-\beta)^2}\right) + (1+\alpha_2)\omega n\frac{c_0}{\eta^{(1-\epsilon)}}$, then the above can be rewritten as :

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \leq R_3\mathbb{E}\|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \hat{\mathbf{X}}^{I_{(t+1)}}\|_F^2 + R_4\mathbb{E}\|\mathbf{X}^{I_{(t+\frac{1}{2})}} - \bar{\mathbf{X}}^{I_{(t+\frac{1}{2})}}\|_F^2 + R_5\eta^2$$

Using the update steps of algorithm given in equations (6) and (10) (given in Section IV):

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \leq R_3\mathbb{E}\left\|\hat{\mathbf{X}}^{I_{(t+1)}} - \mathbf{X}^{I_{(t)}} + \sum_{t'=I_{(t)}}^{I_{(t+1)}-1} \eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}))\right\|_F^2$$
$$+ R_4\mathbb{E}\left\|\bar{\mathbf{X}}^{I_{(t)}} - \mathbf{X}^{I_{(t)}} - \sum_{t'=I_{(t)}}^{I_{(t+1)}-1} \eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}))\left(\frac{\mathbf{1}\mathbf{1}^T}{n} - \mathbf{I}\right)\right\|_F^2 + R_5\eta^2$$

For the same $\alpha_5 > 0$ (from result in Footnote 11) used in proof of Lemma 13, we get:

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \hat{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \leq R_3(1+\alpha_5^{-1})\mathbb{E}\left\|\hat{\mathbf{X}}^{I_{(t+1)}} - \mathbf{X}^{I_{(t)}}\right\|^2 + R_4(1+\alpha_5^{-1})\mathbb{E}\left\|\bar{\mathbf{X}}^{I_{(t)}} - \mathbf{X}^{I_{(t)}}\right\|^2$$
$$+ R_4(1+\alpha_5)\mathbb{E}\left\|\sum_{t'=I_{(t)}}^{I_{(t+1)}-1} \eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}))\left(\frac{\mathbf{1}\mathbf{1}^T}{n} - \mathbf{I}\right)\right\|_F^2$$
$$+ R_3(1+\alpha_5)\mathbb{E}\left\|\sum_{t'=I_{(t)}}^{I_{(t+1)}-1} \eta(\beta\mathbf{V}^{(t')} + \boldsymbol{\nabla F}(\mathbf{X}^{(t')}, \boldsymbol{\xi}^{(t')}))\right\|_F^2 + R_5\eta^2$$

Using $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_2$ to split the third term and then using $\left\|\frac{\mathbf{1}\mathbf{1}^T}{n} - \mathbf{I}\right\| \leq 1$ (from Claim 2 in supplementary material), and further using the bound in (108) for the third and fourth term, the above can be rewritten as:

$$\mathbb{E}\|\mathbf{X}^{I_{(t+1)}} - \bar{\mathbf{X}}^{I_{(t+2)}}\|_F^2 \leq R_3(1+\alpha_5^{-1})\mathbb{E}\left\|\hat{\mathbf{X}}^{I_{(t+1)}} - \mathbf{X}^{I_{(t)}}\right\|^2 + R_4(1+\alpha_5^{-1})\mathbb{E}\left\|\bar{\mathbf{X}}^{I_{(t)}} - \mathbf{X}^{I_{(t)}}\right\|^2 + 2\eta^2 H^2 nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)(1+\alpha_5)(R_3+R_4) + R_5\eta^2$$

Defining $Q_2 = 2H^2nG^2\left(1+\frac{\beta^2}{(1-\beta)^2}\right)(1+\alpha_5)(R_3+R_4) + R_5$ completes the proof of Lemma 14.

APPENDIX I
MEMORY-EFFICIENT VERSION OF SQUARM-SGD

In this section, we provide our memory efficient version of SQuARM-SGD proposed in the main paper in Algorithm 1.

Algorithm 2 Memory-Efficient SQuARM-SGD

Parameters: $G = ([n], E)$, W

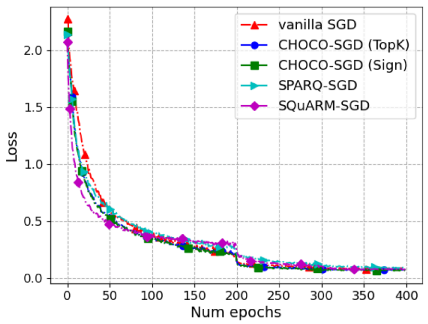
Initialize: For every $i \in [n]$, set arbitrary $\mathbf{x}_i^{(0)} \in \mathbb{R}^d$, $\hat{\mathbf{x}}_i^{(0)} := \mathbf{0}$, $\mathbf{s}_i^{(0)} := \mathbf{0}$, $\mathbf{v}_i^{(-1)} := \mathbf{0}$. Fix the momentum coefficient β, consensus step-size γ, learning rate η, triggering thresholds $\{c_t\}_{t=0}^{T}$, and synchronization set $\mathcal{I}_T$.
for $t = 0$ **to** $T-1$ in parallel for all workers $i \in [n]$ **do**
  Sample $\xi_i^{(t)}$, stochastic gradient $\mathbf{g}_i^{(t)} := \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})$
  $\mathbf{v}_i^{(t)} = \beta\mathbf{v}_i^{(t-1)} + \mathbf{g}_i^{(t)}$
  $\mathbf{x}_i^{(t+\frac{1}{2})} := \mathbf{x}_i^{(t)} - \eta(\beta\mathbf{v}_i^{(t)} + \mathbf{g}_i^{(t)})$
  if $(t+1) \in I_T$ **then**
    for neighbors $j \in \mathcal{N}_i$ **do**
      if $\|\mathbf{x}_i^{(t+\frac{1}{2})} - \hat{\mathbf{x}}_i^{(t)}\|_2^2 > c_t\eta^2$ **then**
        Compute $\mathbf{q}_i^{(t)} := \mathcal{C}(\mathbf{x}_i^{(t+\frac{1}{2})} - \hat{\mathbf{x}}_i^{(t)})$
        Send $\mathbf{q}_i^{(t)}$ to worker j and receive $\mathbf{q}_j^{(t)}$
      else
        Assign $\mathbf{q}_i^{(t)} := 0$
        Send $\mathbf{q}_i^{(t)}$ to worker j and receive $\mathbf{q}_j^{(t)}$
      end if
    end for
    $\hat{\mathbf{x}}_i^{(t+1)} := \mathbf{q}_i^{(t)} + \hat{\mathbf{x}}_j^{(t)}$
    $\mathbf{s}_i^{(t+1)} := \mathbf{s}_i^{(t)} + \sum_{j=1}^{n} w_{ij}\mathbf{q}_j^{(t)}$
    $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t+\frac{1}{2})} + \gamma\left(\hat{\mathbf{s}}_i^{(t+1)} - \hat{\mathbf{x}}_i^{(t+1)}\right)$
  else
    $\hat{\mathbf{x}}_i^{(t+1)} = \hat{\mathbf{x}}_i^{(t)}$, $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t+\frac{1}{2})}$, $\mathbf{s}_i^{(t+1)} = \mathbf{s}_i^{(t)}$
  end if
end for

The parameter $\mathbf{s}_i^{(t)}$ for $i \in [n]$ stores the weighted sum of all neighbor copies which is then used in the consensus step. Thus, the requirement for storing copies of all neighbors at a node as in algorithm given in main paper is relaxed.

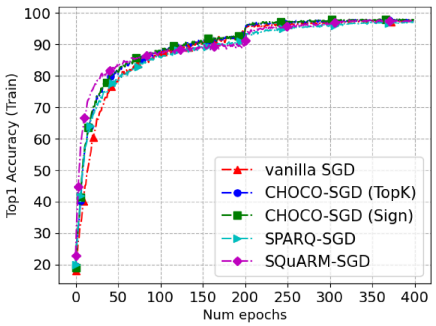
APPENDIX J
ADDITIONAL EXPERIMENTS

In this section, we provide additional experiments for comparison of schemes when training a ResNet-20 model on the CIFAR-10 dataset, with the same setting as Section VI in the main paper.

A. Training performace



(a) Training loss vs epochs for all schemes.

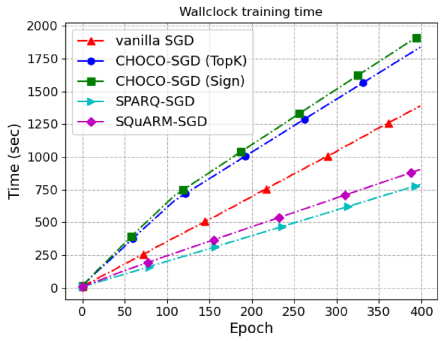


(b) Training accuracy vs epoch for all schemes.

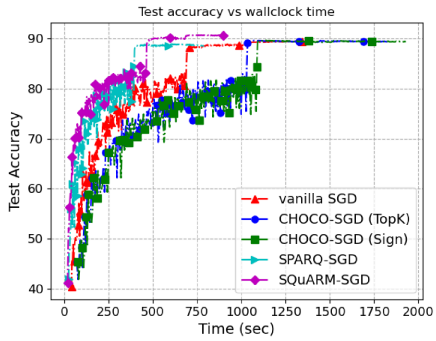
Fig. 3 Training metrics for different schemes.

Figure 3 shows the training loss and training accuracy performance of all the schemes. We observe that each scheme is able to train the ResNet-20 model well over the CIFAR-10 dataset.

B. Wall clock comparison



(a) Wall-clock training time logged at each epoch.



(b) Test accuracy vs wall-clock time.

Fig. 4 Comparing performance of schemes with wall-clock training time.

Figure 4a shows the wall-clock time for training the ResNet-20 model for all the schemes logged in at each epoch. It can be seen that performing the encoding/decoding process for CHOCO-SGD (Sign/TopK) [21] can be expensive, and takes more time than vanilla SGD. For SPARQ-SGD and SQuARM-SGD, we consider 10 local iterations, and thus the nodes only need to perform the encoding decoding process once in every 10 iterations as compared to each iteartion in vanilla SGD or CHOCO-SGD. The time take for SQuARM-SGD is a bit higher than SPARQ-SGD on account on performing more computation with the momentum updates.
Figure 4b shows the test error performance as a function of the wall clock time elapsed during training. It can be seen that on account of using momentum and local iterations, SQuARM-SGD achieves a higher test performance while taking about $0.5\times$ the time compared to CHOCO-SGD for training, and about $0.75\times$ the time compared to vanilla-SGD.

C. Effect of local iterations (H) on SQuARM-SGD

We consider the same setting as in our main paper, and compare the performance of SQuARM-SGD for $H \in \{5, 20, 70, 100\}$. We observe that the proposed scheme works well for reasonable values of H, which taking large H can slightly hurt the performance due to very infrequent information exchange among clients.

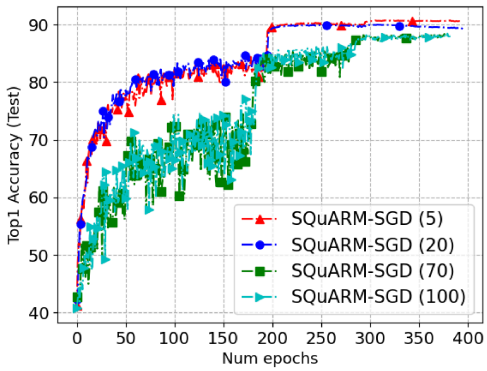


Fig. 5 Comparison of Test accuracies for different values of H for SQuARM-SGD.